Improving Equitable Measurement and Reporting in NAEP

Gerunda B. Hughes
Howard University

June 2023
Commissioned by the NAEP Validity Studies (NVS) Panel

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Keena Arbuthnot
Louisiana State University

Peter Behuniak
Criterion Consulting, LLC

Derek Briggs
University of Colorado Boulder

Jack Buckley
American Institutes for Research

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Gerunda Hughes
Howard University

James Pellegrino
University of Illinois at Chicago

Gary Phillips
Cambium Assessment

Jennifer Randall
University of Michigan

Akisha Osei Sarfo
Council of the Great City Schools

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
College Board

Project Director:

Sami Kitmitto
American Institutes for Research

Project Officer:

Grady Wilburn
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS) Panel
American Institutes for Research
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
Email: naepvaliditystudies@air.org

CONTENTS

ABOUT NAEP

The National Assessment of Educational Progress (NAEP), also known as "the Nation's Report Card," is the only nationally representative assessment of what U.S. students know and can do in various subject areas, such as reading, mathematics, science, writing, civics, geography, technology, and engineering literacy. NAEP has provided information about how students are performing academically since 1969, and many policymakers, researchers, and the media view NAEP as the "gold standard" for monitoring and documenting what fourth, eighth, and twelfth graders in the nation, states, and select urban districts know and can do in various subject areas. The purpose of NAEP is to both "lead and reflect" what is taught in the states. In addition to the Main NAEP assessments, most often at Grades 4 and 8 and occasionally at Grade 12, the NAEP long-term trend (LTT) assessments give information on changes in the basic achievement of America's youth. LTT assessments are administered nationally and report student performance for students ages 9, 13, and 17 in reading and mathematics.

The NAEP Law

NAEP is congressionally mandated and administered by the National Center for Education Statistics (NCES) within the U.S. Department of Education and the Institute of Education Sciences. According to the National Assessment of Educational Progress Authorization Act of 2002 (Pub. L. 107-279, Section 303), also known as "the NAEP Law," the purpose of NAEP is "… to provide, in a timely manner, a *fair and accurate measurement of student academic achievement and reporting of trends* [emphasis added] in such achievement in reading, mathematics, and other subject matter as specified in this section" (20 U.S. Code § 9622(b)(1)). The Act charges the Commissioner of Education Statistics to carry out several duties related to NAEP, of which the following are notable:

- Use widely acceptable and professional random sampling processes that produce representative data on a national and regional basis.
- Conduct a national assessment and collect and report assessment data on the status of achievement as well as achievement data trends and do this in a valid and reliable manner for student groups in both public and private schools.
- Include information on special student groups, including whenever feasible information collected, cross tabulated, compared, and reported by race, ethnicity, socioeconomic status, gender, disability, and limited English proficiency.

It is worth repeating that the NAEP Law states that the purpose of NAEP is to provide a "*fair and accurate measurement of student academic achievement"* [emphasis added] (20 U.S. Code § 9622(b)(1)). According to *Educational Assessments in the COVID-19 Era and Beyond* (National Academy of Education, 2021), an equitable educational assessment system is one that is fair and accurate. Fair assessments are sensitive to the characteristics of different groups being assessed and, thereby, where appropriate, reflect flexibility in the design and administration of the assessment and the reporting of the assessment results. This notion of fairness also is in harmony with how fairness is described and discussed in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

NAEP has a long-standing reputation for providing valid and reliable information about what American students know and what they can do at specific grade and age levels; therefore, given the diversity that exists among American students, I pose the following question for thoughtful consideration:

- Do current practices and methodologies used by NAEP (e.g., design, development, administration, scoring) mostly assume that American students are part of a "melting pot" socially and culturally or do they acknowledge and reflect that America has become more socially and culturally diverse since the inception of NAEP?

Surely, as social and culture structures change in America, testing and assessment—and specifically the components of measurement and reporting—will need to adapt to this ever-evolving reality (Basterra et al., 2011; Bennett, 2022; Randall et al., 2022). To that end, Bennett (2023) proffers a theory of socioculturally responsive assessment in which an assessment, *given its purpose and context*, can be designed to connect to the cultural identities, backgrounds, and lived experiences of a diverse group of individuals, and especially those from traditionally underserved groups. This theory has the potential to provide a "theory of action" for NAEP. Consequently, while the title of this paper acknowledges that elements of equitable measurement and reporting currently exist in NAEP to some degree, it also recognizes or affirms that there is room for improvement.

A Historical Perspective on Validity, Fairness, and Equity in Large-Scale Assessments

The issue of fairness and equity in educational and psychological assessment—which includes the development, administration, scoring, interpretations, and uses of assessment results—has been the subject of civil discussions and heated debates for decades (Bennett, 2021; Cleary, 1968; Cole, 1973; Darlington, 1971; Davis, 1948; Delandshere & Petrosky, 1998; Gordon, 1995; Gordon, 1999; Jensen, 1980; Johnson, 1979; Linn, 1973; Schmidt & Hunter, 1974; Sireci, 2020, 2021; Solano-Flores & Trumbull, 2003). In fact, questions and concerns about fairness in assessment have been raised by individuals both inside and outside the assessment community and by individuals in the classroom, in the courthouse, and even on Capitol Hill. Most notably, issues about fairness and equity in assessment have been raised by members of communities in which fairness in testing and assessment seem illusive, even though there is a plethora of research studies on fairness within the field (Arbuthnot, 2011, 2015; Ford, 2004, 2007; Hilliard, 1991; Hood, 1998; Lee, 1998; Williams, 1970). Although measurable progress might seem slow to some individuals in the field, advocates for fairness and equity in assessment continue to raise their voices and use the power of research and professional activism within the fields of assessment and measurement. Also, provisions of landmark legislation and the lessons learned from the results of litigation have helped advance fairness in testing and assessment and in the broader educational system (see the appendix). Still, there is so much more that can and should be accomplished.

Across time, researchers in the testing, assessment, and measurement field, and especially researchers from marginalized groups, have challenged many of the assumptions, policies, and practices that support the testing, assessment, and measurement enterprise. These assumptions, policies, and practices have been challenged because they are perceived and have been shown through research to be roadblocks to achieving equity and fairness for

some demographic groups (Gould, 1996; Helms, 1992; Johnson, 1980; Meredith, 1993; Royer & Carlo, 1993; Solano-Flores & Nelson-Barber, 2001). These roadblocks include but are not limited to (a) standardization, (b) assumptions of construct equivalence across culturally different test takers, (c) use of item statistics for test construction that may disadvantage low scorers; and (d) the reporting of test results in ways that highlights the superior performance of White test takers to that of Black or Hispanic test takers but obscures the superior performances of Asian/Pacific Islanders compared with White test takers. Many of these roadblocks have remained in place for decades, even though they are regularly examined and reported on by researchers in the field. Here are some examples.

The Promise and Peril of Standardization Throughout the Testing Process

The standardization of various aspects of the development, administration, and scoring of educational tests is necessary to make meaningful interpretations or comparisons of the performance of test takers. Most educational tests are standardized on the same content (e.g., same domains of subject matter), under some of the same testing conditions (e.g., same testing window), and with the same scoring protocols (e.g., the same rubric or rules). Tests designed for credentialing, including licensure and certification, for example, can be expected to have a high level of standardization at various steps in the testing process. A reading comprehension test for eight graders that requires all test-takers to read and answer questions based on the same reading passage, however, could allow for more flexibility. Although one purpose of standardization is to "level the playing field" for all test takers and promote fairness in testing, some types of standardized practices can, in fact, have the opposite effect and contribute to unfairness for some test takers. All test-takers may not have the same level of interest or engagement in the passage selected by the test developer which could adversely affect the measurement of their reading comprehension. Sireci (2021) noted that overly rigid testing procedures can impede the accurate measurement of students' proficiencies, distort test score interpretations, and lead to inappropriate uses of test results.

Valid interpretations of students' test scores require understanding how their personal or group characteristics may interact with the standardized testing conditions. For example, "translanguaging," which is a departure from standardized testing conditions, refers to the "flexible use of linguistic resources [during testing] that characterizes bilinguals in their attempt to make sense of their bilingual worlds" (Gandara & Randall, 2019, p. 63). Thus, solely relying on monolingual expectations and procedures for test administration may limit the opportunity for bilinguals to access the fluid and strategic uses of multilingual resources that they possess (Gándara, 2017). The idea that all test takers respond to "standardization" in the same way does not acknowledge the multiplicity of ways of knowing and functioning, given the diversity of the world in which we live (Arbuthnot, 2011, 2020).

Assumptions of Construct Equivalence and Universality Across Culturally and Linguistically Different Subgroups of Test Takers

Dixon-Román and Gergen (2013) posed the following question: "Is a test of reading ability simply this, or is it a test of home environment, economic privilege, or parental influence?" (p. 19). The authors raised this question because, in practice, measurement sciences have continuously supported "assumptions of universality and have made strong assumptions about the minds and actions of those being measured" (p. 16). In doing so, it has privileged the dominant culture's judgments about those who are measured at the cost of acknowledging the multiplicity of social, cultural, and language variants that exist in society

(Solano-Flores & Trumbull, 2003). Dixon-Román and Gordon (2012) noted that education is essentially a social process that reflects the needs, values, and expectations of local populations; measurement modeling, on the other hand, as currently practiced, is at odds with the sociocultural relational processes of education.

Randall (2021) noted that students do not experience the world, including schooling, in ways that are context free; so why is there an assumption of universality when they experience tests and assessments? Sociocultural contexts related to differences in race/ethnicity, language, family economic status, parental levels of education, geography, and religion—to name a few—individually and in combination have a profound effect on how intended constructs are perceived by students from diverse backgrounds and how they are operationalized in their responses to test items. Thus, "construct definition," and, more precisely, "construct perception and operationalization" are central to any validity argument about the interpretations and uses of test scores, especially for marginalized groups.

To account for the various ways in which test takers legitimately interact with test items and assist a test developer with the important job of defining constructs for multiple ways of knowing and understanding phenomena, Randall (2021) developed "a heuristic for a justice-oriented, antiracist approach to construct definition and representation" (p. 7). The heuristic comprises a series of questions the test developer could ask to enhance self-awareness about the process of construct definition. For example,

- "What racial identity [knowing or unknowingly] am I [the test developer] bringing into the construct definition process?" (positionality)
- "What are the sociocultural and racial contexts, resources, hopes, dispositions, experiences, and backgrounds of stakeholders?" (people/places)
- "If one language, literacy, culture, or way of constructing knowledge is privileged over another, how have you [the test developer] attempted to disrupt this imbalance of power?" (power)

In addition, to accompany the heuristic for construct definition, Randall et al. (2022) developed a heuristic for a justice-oriented, antiracist approach for building a solid validity argument. This heuristic includes questions such as the following:

- "Have a wide range of interpretations been considered that acknowledge the different ways of knowing?" (response processes)
- "Do the test/assessment results serve to further marginalize already minoritized groups?" (consequences)

Concerns about the lack of acknowledgment of the sociocultural influences on test performance by test developers and users of test results did not begin at the end of the 20th century and beginning of the 21st century, as these references might suggest. During the 1970s, for example, many articles related to "culture-free" and "culture-fair" tests appeared in the research literature (Barnes, 1972; Williams, 1970, 1972). In most cases, culture free/fair tests developed at that time were purported to be intelligence tests; however, many of them failed to pass the "validity test" of being free of cultural influences or fair in terms

of their accurate interpretations of the intelligence of children who may be marginalized (Hilliard, 1979a, 1979b).

To expose the mythology of a culture free/fair test, Barnes (1972) and Williams (1970) proposed the development of "culture-specific tests." The purpose of such tests was to determine a test taker's ability to "think in terms of his own culture and environment" (Barnes, 1972, p. 6), which the *Stanford-Binet Intelligence Scales* or the *Wechsler Intelligence Scale for Children* does for White test takers. The implication is this: If a child can demonstrate capabilities of conceptual thinking in his own cultural context, he also can demonstrate and master concepts in the school curriculum that require similar capabilities. Consequently, in 1972, Robert Williams, an African American psychologist, developed a culture-specific test, the *Black Intelligence Test of Cultural Hegemony* (BITCH-100). The BITCH-100 was designed to include content material familiar to African Americans: their language and familiar idioms. As expected, when the BITCH-100 was administered to samples of Black and White 16–18-year-olds, students in the Black sample scored significantly higher with less variance (M = 87.0; SD = 6.97) than students in the White sample (M = 57.0; SD = 16.2). Williams (1972) did not use the relatively poor performance of the White students to label them as significantly inferior to the Black students. Rather, he suggested that the results for the White sample would be useful as a measure of the level of awareness and familiarity of a White person with the Black experience. Consequently, the results could be used, for example, as a predictor of the level of empathy a White person is likely to display who cares for Black patients or teaches Black children or the White law enforcement personnel (e.g., police officers) who service Black neighborhoods or interact with Black activists (e.g., protesters for social justice).

Item Statistics in Test Construction Disadvantage Low Scorers Who Are Likely From Marginalized Groups

Making fair judgments about performance on tests can be very challenging. Judgments should not be made solely on the basis of intuition or what something appears to be, and they should not be made using strategies or practices that marginalize groups of test takers with certain defined social, cultural, or linguistic characteristics. When tests are used to make such judgments, conducting comprehensive item and test analyses – statistical and judgmental - are the best practices for providing information about the quality of individual items and the overall test. Statistical item analyses yield two item statistics of interest: item difficulty and item discrimination. These statistics can be used along with judgmental evaluations to revise items or make decisions about whether an item will be included in the operational version of the test. The decisions that test developers make about inclusion or exclusion of an item during test construction can (a) provide counterintuitive interpretations for educators, specifically teachers, about what the value of the item statistics mean in a standards-based educational system; (b) have a negative effect on the performance of some test takers, especially low scorers; and (c) lead to invalid conclusions about what these test takers really know and can do (Hilliard, 1991).

Item difficulty (p) is the percentage of test takers who answer an item correctly. To compute item difficulty, divide the number of test takers who answered the item correctly by the total number of test takers. The larger the value of p, the easier the item; conversely, the smaller the value of p, the more difficult the item. If item A has an item difficulty of $p = 1$, then every test taker answered item A correctly. If item B has an item difficulty of $p = 0$, then

every test taker answered item B incorrectly. Psychometricians and measurement scientists interpret these values to mean that item A and item B do not contribute to measuring individual differences among test takers and therefore are useless for discriminating among different groups of test takers. Additionally, they conclude that item A and item B do not provide information for constructing a scale because there is no variation in the item scores among test takers. That may be true, psychometrically, but for a teacher, the *p*-value of item A tells her that whatever concept item A is assessing, all students demonstrated their knowledge (or even mastery) of that concept. That kind of performance calls for a celebration for both teachers and students in a standards-based, accountability educational setting. In addition, the *p*-value for item B signals that the teaching and learning of that concept needs to be revisited.

After calculating the *p*-values for items A to Z, the item analyses may reveal that the test has other items on the test with $p = 1$. If most items, C to Z, have *p*-values that do not equal 1 or 0 but are judged to be good discriminators among groups of test takers, then items with $p = 1$ may be eliminated from the operational test. Yet, these are the very items that low scorers answered correctly! Therefore, it begs the question: "Were low scorers given the opportunity to show all they know and can do?"

Robert Williams described the challenges he had with using item difficulty and item discrimination to make inferences about the validity of some test results:

> As mentioned earlier, so-called easy test items, as a rule, do not discriminate. They are therefore discarded. In the context of this paper, the definition of easiness or difficulty is relative. What is easy for one group is difficult for another group. Proponents of testing claim that a negatively skewed test is useless because the items do not measure individual differences among the more able subjects of the group. They suggest including more difficult items in order to [en]sure a normal distribution. The writer would have to question this search for individual differences particularly in a society that is ostensibly based on equality. Anastasia (1968) claims that a test is "excess baggage" if everyone passes every item. If everyone passes every item, one might have a great deal of information about the group as a whole. In disagreement with Anastasia, such a test might prove more reliable than one whose items are scaled according to difficulty. The test may be more valid than one which reflects individual differences. (Williams, 1972, pp. 13–14)

I understand the challenges Williams had with using item difficulty and item discrimination to make inferences about the validity of some test results – especially in the context of large-scale testing and accountability. It seems more appropriate to administer tests with items that discriminate and measure individual differences *during instruction* so that those differences in performance can be eliminated. Furthermore, tests used for accountability or monitoring purposes should be constructed in ways that allow all students to fully demonstrate what they know and can do. To design, administer, and score them otherwise seems highly counterintuitive and unfair.

Reporting Test Results That Highlight the Superior Performance of White Test Takers Compared With Black and Hispanic Test Takers but Obscure the Superior Performance of Asian/Pacific Islanders Compared With White Test Takers

Designing tests to measure individual and group differences is like a two-edged sword that cuts both ways. It can reveal strengths or weaknesses for any group depending on how the test data are analyzed and reported. In the United States, comparisons of test performance based on racial/ethnic groups are the focus of vast numbers of research reports and journal articles, books, dissertations, newsletters, and websites. The NAEP Law provides for the comparison of information by race and ethnicity in Section 303, Part G of Public Law 107-279. The most popular comparison of test performance by race/ethnicity reported widely is performance between White and Black or African American individuals. I wrote a review of the book *The Black-White Test Score Gap* (Jencks & Phillips, 1998). The first sentence in the first paragraph of the first chapter makes a profound point: "African Americans score lower than European Americans on vocabulary, reading, and math tests, as well as on tests that claim to measure scholastic aptitude and intelligence" (Jencks & Phillips, 1998, p. 1). If a reader does not encounter another sentence in the book, what message would the reader perceive the authors are communicating? What message was received? I often have wondered whether the design, administration, and scoring of standardized tests along with the perpetual inequities in all aspects of teaching, learning, and living are intentional in producing the evidence in this quote.

Interestingly, what frequently is not presented in oral presentations or published in journal articles, books, and technical reports is the fact that on NAEP, Asian/Pacific Islander schoolchildren in the United States outperformed White schoolchildren in reading and mathematics, in both fourth and eighth grade, for all years that the Main NAEP was administered from 2011 to 2017. This begs the question: "Why hasn't this information about the achievement gap between Asian Americans and European Americans been discussed openly and reported widely?"

A Call for Action

Therefore, given these considerations of the entrenched roadblocks to fairness and equity in the testing, assessment, and measurement field, there continues to be a collective and bold call to action among impacted members in the field, who bring a unique, personal, and professional perspective, along with their supportive colleagues, to address and resolve concerns related to fairness and equity in testing (Gordon Commission, 2013).

A Focus on the Validity of Assessment Results: Interpretations and Uses

Two important issues related to fairness and equity in testing and assessment are the accuracy of the interpretations and the appropriateness of the uses of the assessment results, given the intended purpose of the assessment (Davis, 1948; Gordon, 1995; Jensen, 1980; Linn, 1973). Together, these two measures speak to the validity of the assessment results for a given purpose. The accuracy of the interpretations of the assessment results is important because many variables relate to performance on a test, and it is not always easy to identify where in the testing (or teaching and learning) process alleged inequities occur or why they occur. For example, the individual characteristics of test takers, such as race, ethnicity,

gender, culture, language, disability, or socioeconomic status, may interact with the testing process at any point in ways that might interfere with a test taker being denied an opportunity to demonstrate their full potential or maximum performance. The interactions between test takers and the testing process produce test scores. Test scores are interpreted by educational personnel who give meaning to the scores. Inaccurate interpretations of test scores may lead to inappropriate uses of the test results, which lead to unintended negative consequences for individuals, groups, or systems.

Consequently, in recent years, with the proliferation of a "test-based educational system," the development and administration of tests, as well as the interpretations and uses of test results, especially for so-called accountability purposes, have come under intense public scrutiny. Specifically, the call for more in-depth explanations about the differential performance of distinct groups defined by race, ethnicity, gender, language, or disability has led to renewed interest in and scholarship about the meaning and the role of fairness and equity in assessment.

It also is important to acknowledge that differential performance of diverse groups on tests does not necessarily constitute unfairness. Even though perennial achievement gaps continue to exist almost ad infinitum between students of different racial, ethnic, cultural, and socioeconomic groups, there is a desire among researchers in the impacted communities to understand and separate the role of inequities in the educational system from the role of inequities in the assessment and measurement processes that may explain some of the achievement gap. According to the *Standards for Educational and Psychological Testing*, "Regardless of the purpose of testing, the goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure" (American Educational Research Association et al., 2014, p. 51). Furthermore, ETS, formerly known as the Educational Testing Service, in its *Standards for Quality and Fairness* provides the following statement about fairness for guiding test construction: "Fairness is the extent to which the inferences made on the basis of test scores are valid for different groups of test takers" (Educational Testing Service, 2014, p. 19). Thus, based on the stated positions in documents such as the *Standards for Educational and Psychological Testing* and *Standards for Quality and Fairness* as well as the provisions of the NAEP Law, equitable measurement and reporting in NAEP can and should be improved.

Validity, Fairness, and Equity in Testing and Assessment

Validity

Validity is the most fundamental consideration in developing and evaluating tests. It is the degree to which evidence and theory support the interpretations of test scores for the proposed uses of those tests. But the validity of interpretations and uses of scores does not exist within a vacuum; it is explicitly connected to the construct(s) the test is intended to measure as well as the interactions of the test-takers' characteristics and the test. Thus, evaluating tests, and by extension validity, involves collecting relevant evidence to provide a sound scientific basis for the proposed score interpretation(s) and use(s) related to the construct.

Cizek (2020) noted that "a primary element in the definition of validity is the role of variation—variation in examinees' (and by extension subgroups') standing on the construct of interest, and variation in their observed test performances" (p. 98). Consequently, validity

represents the extent to which variance in examinees' standing on the construct is reflected (as in a one-to-one correspondence) in their test scores. There are essentially two reasons why examinees' standing on a construct vary: (a) the examinees' standing on the construct is caused by a single, perhaps composite factor (that is, the construct is unidimensional as it interacts with the examinees); or (b) the examinees' standing on the construct varies because of a variety of specified factors unrelated to the construct definition (i.e., the construct is multidimensional as it interacts with the examinees). Therefore, the primary goal of a validity study is to identify and determine the extent to which specific factors account for the variation in scores among examinees or subgroups. Cizek (2020) noted that "construct-relevant variation" exists in the "*ideal*" [italics mine] validity situation when all of the variance in examinees' observed test scores corresponds to variance in their standing on the intended construct. However, if other factors deemed not related to the intended construct affect examinees' test scores, then, psychometrically speaking, these factors are said to be sources of construct-irrelevant variation. Because the ideal validity situation is "highly improbable," that is, when there is a high probability that variation in examinees' test scores is caused by factors deemed not related to the intended construct, then test developers and test score users have reason to be concerned about the accuracy of the interpretations and inferences of test scores for all examinees and for marginalized subgroups, in particular. In other words, validity becomes an issue. Cizek (2020) noted as follows:

> [T]he desire of a test developer [is] that scores yielded by an instrument can be interpreted by consumers of that information as intended . . . [S]ources of construct-irrelevant variance are particularly troublesome. Construct-irrelevant variance contributes to variation . . . in a way that makes it *appear* that examinees [or subgroups] differ on the construct of interest, but in fact the observed difference may be illusory, leading to inaccurate interpretations of those scores. (p. 99)

According to this statement, one might conclude that the best way to improve validity is to reduce the influence and effect of construct-irrelevant factors. Zieky (2016) noted that only construct-relevant sources of variance enhance validity; all other sources of variance weaken validity. Furthermore, if construct-irrelevant sources of variance, such race/ethnicity, language, culture, economic status, or parental level of education, are correlated with group membership, then validity and fairness are weakened. But what if there is a way of looking at sources of so-called construct-irrelevant variance, not as irrelevant but as central and relevant to how the intended construct is perceived and operationalized in responding to test items by culturally diverse subgroups?

On Cultural Validity in Testing and Assessment

Tests are cultural artifacts and viewing them as such help developers of tests and users of test results appreciate how students from diverse cultural backgrounds interact with tests—how they interpret test items and respond to them, which may be different from the intended knowledge or skills being assessed (Durán, 2011; Solano-Flores, 2011). This is a matter of validity because, according to "conventional wisdom," student performance on a test should not be influenced by factors other than those that the test is intended to measure (Messick, 1995). Yet, even though the influence of culture and language on test performance is acknowledged as sources of construct-irrelevant variance and measurement error, current testing practices address culture and language as a threat to validity rather than the essence of validity. Solano-Flores and Nelson-Barber (2001) proposed the concept of cultural validity as follows:

> [T]he effectiveness with which . . . assessment addresses the socio-cultural influences that shape student thinking and the ways in which students make sense of . . . items and respond to them. These socio-cultural influences include the sets of values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in the students' cultural backgrounds, and the socioeconomic conditions prevailing in their cultural groups. (p. 555)

Solano-Flores and Nelson-Barber (2001) acknowledged that the cultural factors that influence ways of thinking, learning, and knowing also are present in tests and test taking; these influences do not disappear. Hence, culture should not be treated as an incidental or nonexistent factor in the interpretations and uses of test results; it should be treated as a phenomenon on which the foundations of test design, development, administration, scoring, and reporting are based. Furthermore, Solano-Flores and Nelson-Barber (2001) argued that both test developers and test users should examine cultural validity with the (a) same level of vigor and attention they use when they examine other forms of validity and (b) use more effective tools and techniques for testing construct and measurement invariance.

The *Standards for Educational and Psychological Testing* state as follows: "It is the interpretations of the test scores for the intended uses that are evaluated, not the test itself" (American Educational Research Association et al., 2014). However, it also can be argued that features of the test (i.e., test design, administration, and scoring) interact with the characteristics embodied in the diversity of the population of test takers that produces the test scores. The test scores are then reported and shared with a variety of audiences. *An Agenda for NAEP Validity Research* noted as follows: "Validity is the extent to which the messages in NAEP reports accurately communicate the state of educational progress in America to educators, policymakers, and the public" (Stancavage et al., 2003, p. i). The importance of valid interpretations of test scores cannot be overstated because many test-based performance inferences or decisions are made by educators, policymakers, and the public about individuals, groups, or systems that may be judged as unfair or that may, for some entities, have serious economic, psychological, social, or educational consequences.

Fairness

The 2014 edition of the *Standards for Educational and Psychological Testing* devotes an entire chapter to fairness in testing, in which the authors note that their focus would be "to delineate the aspects of tests, testing, and test use that relate to fairness which are the responsibility of those who develop, use, and interpret the results of tests, *and upon which there is general professional and technical agreement"* [italics mine] (American Educational Research Association et al., 2014). The caveat at the end may leave loopholes for determining whether certain aspects or features of the test or the testing process are deemed or determined to be unfair.

The current edition of the *Standards* also acknowledges that fairness has no single technical mechanical meaning; there is no "fairness scale" on which a test is evaluated. However, the article asserts that fairness is a fundamental validity issue that requires attention throughout all stages of test design, administration, scoring, interpretation of results, reporting, and use. The authors of the *Standards* also note that fairness to all individuals in the intended population of test takers is an overriding foundational concern, and common principles apply in responding to test-taker characteristics that could interfere with the validity of test score interpretations. Such characteristics that may interfere or interact with the testing process include but are not limited to

those defined by race, ethnicity, gender, culture, language, age, disability, or socioeconomic status.

Accessibility to the Test

When construct-irrelevant characteristics interact with test performance for some test takers, one solution may be to invoke different or equitable strategies that level the playing field in the testing context. For example, the unobstructed opportunity to demonstrate standing on the construct(s) being measured is referred in the *Standards* as "accessibility." Individuals with physical or cognitive disabilities may be disadvantaged because of visual impairment or may require additional testing time; test takers classified as limited English proficient may require test translations or language simplifications to maximally demonstrate their standing on the test. Test takers classified in these two categories are "protected" and provided "accommodations" to access the construct(s) being measured; otherwise, the validity of the score interpretations for the intended uses for individuals in these protected groups is threatened. Other examples of more commonly used accommodations include but are not limited to (a) administering tests in a separate context (e.g., one-on-one test administrator); (b) providing oral accommodations (e.g., read test directions aloud); or (c) presenting the test content in different formats (e.g., adjusting the font size or using a paper and pencil version versus an electronic device). The availability and use of these accommodations during test administration allow a test taker to participate in the test more equitably. For protected groups, access to these types of accommodations is a legal requirement in some testing contexts (refer to Individuals with Disabilities Education Act of 2004; No Child Left Behind Act of 2001). These legal protections promote fairness in testing (Abedi & Ewers, 2013; Faulkner-Bond & Soland, 2020; Sireci et al., 2018). Currently, no such legal protection with respect to accessibility is currently available for test takers whose culture or socioeconomic status might interact with the testing process in ways that obstruct their opportunity to demonstrate maximally their standing on the construct(s) being measured.

In summary, the *Standards* "interprets fairness as responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses" (American Educational Research Association et al., 2014, p. 50). From the perspective of the *Standards*, a test is fair if it reflects the same construct(s) for all test takers, and scores from the test have the same meaning for all individuals in the intended population. Accordingly, a fair test does not create an advantage or a disadvantage for some individuals because of characteristics irrelevant to the intended construct.

Equity

In his 1995 seminal article, *Toward an Equitable System of Educational Assessment*, Edmund W. Gordon explained that equity speaks to differences, especially the differential or unequal distribution of resources and inputs to meet a specific need or provide an opportunity to level the playing field. Testing accommodations are an example of equity in testing because they represent adjustments in the standard testing administrations that level the playing field for access to the construct(s) being measured and promote the valid interpretations of scores. Relatedly, Gordon (1995) posited that equity "speaks to and references fairness and social justice" (p. 363).

Equality, on the other hand, speaks to sameness or no difference. In the testing context, one might liken equality to standardization. According to the *Standards*, for decades of testing,

standardization was (and is) viewed by many measurement professionals as a fundamental principle for ensuring that all test takers have the same opportunity to demonstrate their standing on the construct(s) being measured. Operationally, standardization requires that all groups of test takers (except those protected by federal laws) adhere to a uniform set of testing features related to test design, administration, scoring, analyses, and interpretation of the results. Standardization does not consider the diversity of characteristics among test takers that may interact differentially at any stage along the testing process.

In most large-scale testing contexts, standardization assumes comparability or sameness of characteristics for all test takers at a macro level (e.g., all test takers are in the same grade) but ignores the diversity and heterogeneity of characteristics at a micro level that explain how individuals or groups interact with the testing context (Berman, et al., 2020; Linn & Harnisch, 1981). This assumption of comparability of the intended population of test takers influences test design, administration, analysis, and the subsequent interpretations and uses of the test results. Given the way it is currently implemented in testing and assessment, "standardization is the epitome of unfairness" (Pellegrino, 2021b, 1:40:40).

The goal of employing equitable practices in the testing process, where appropriate, is not to ignore or dismiss appropriate and legitimate reasons for implementing standardization in testing and assessment. Rather, as Sireci (2020) noted, the goal is to level the playing field for all test takers by reconceptualizing standardization in ways that build in flexibility, rather than rigidity, into the testing process. Rigid adherence to standardized testing features that are external to the requirements for accommodations may limit some students' access to the test and result in unintended errors in the measurement of their achievement.

The foregoing discussion reveals the inherent relationships among validity, fairness, and equity. There is a "transitive relationship" among them such that in the testing and assessment context, validity is related to fairness; fairness is related to equity; and, consequently, it can be concluded that validity is related to equity. There are models in the research literature, professional organizations, the courts, and legislatures for examining, promoting, and improving the validity and fairness of testing and assessment. These models can identify ways to improve equity in testing and assessment.

I will now review and explore ways to improve equity in testing and assessment in NAEP, in particular, through an examination of the conception, birth, and developmental evolution of NAEP; governance of NAEP; and efforts to evaluate the validity and utility of NAEP. I will also present a call for action through professional activism directed specifically at policymakers, test developers and measurement specialists, and users of assessment results. Finally, I conclude with recommendations for improving equitable measurement and reporting in large-scale assessments, in general, and in the National Assessment of Educational Progress (NAEP), in particular.

TOWARD EQUITABLE ASSESSMENT, MEASUREMENT, AND REPORTING IN NAEP

NAEP began, in part, as a response to the recommendations of a committee appointed by the U.S. Commissioner of Education, Francis Keppel. In 1963, Keppel asked Ralph Tyler, director of the Center for Advanced Study in the Behavioral Sciences at Stanford University, to chair a committee to consider the feasibility of developing a plan for the periodic national assessment of student learning. The committee also was charged with exploring options for assessing the condition and progress of American education. Although both Keppel and Tyler had compatible interests in creating a national assessment to evaluate student learning, they had somewhat different visions about how to conceive it. According to Jones and Olkin (2004), in *The Nation's Report Card: Evolution and Perspectives*, Tyler believed:

> (a) "that commonly used standardized achievement tests did not provide a valid measure of what children have learned but were designed to rank students"; (b) "the purpose of standardized tests was to identify individual differences in achievement, not to measure individuals' learning"; and (c) "the manner in which standardized tests were scored and reported was not a meaningful way to score and report the achievement of a community" (p. 26).

Keppel, on the other hand, wanted to have national data that would meet the intent of the legislation that created the Department of Education. Keppel believed that standardized testing could provide that kind of data.

After a few years of bold commitment to the goal of designing a national assessment, Tyler (1966) and the committee recommended the development of a battery of tests using the highest psychometric standards and with the consent and blessing of those who would use the data and information at the local, state, and national levels. NAEP was conceived to provide that information and monitor the progress of American education.

To meet this challenge, the first NAEP was administered in 1969 and produced nationally representative assessment data for the content areas of citizenship, science, and writing for 17-year-old students still enrolled in school. In 1970 and 1972, NAEP began testing nationally representative samples of students ages 9, 13, and 17 in mathematics and reading. Test questions, items, or tasks were developed that represented a one-to-one correspondence to the learning objectives of school curricula. Assessment results were tabulated by age and by demographic groups within age but never by state, school district, school, or individual. Assessment results also were reported to show the estimated percentage of the population or subpopulation that answered each item or task correctly.

This design for NAEP reflected the political and social realities of the mid-1960s. At that time, state and local leaders feared federal erosion of their autonomy to develop their own curricular content and performance objectives; therefore, there was wide skepticism about any assessment project that might lead to a national curriculum. Hence, NAEP's early design featured the administration of items and tasks to students defined by age instead of specific grades; and assessment results were reported by items or tasks instead of broad knowledge and skill domains that could be construed as representing some type of nationally prescribed curriculum.

Civil Rights, ESEA (1965), and NAEP

In the 1950s and 1960s, the educational landscape began to change in large part to the unanimous Supreme Court decision in *Brown v. Board of Education of Topeka* (1954), the civil rights movement, and the resulting federal legislation regarding schooling and housing. There was a dramatic increase in the racial and ethnic diversity of the school-aged population and a heightened commitment to providing educational opportunity for all. President Lyndon Johnson laid out his agenda to declare a "war on poverty" and urged Congress to enact the most comprehensive federal education legislation to date.

The goal of the Elementary and Secondary Education Act (ESEA) of 1965, which was part of President Johnson's war on poverty, was to address the problem of inequality in education related to children's or their families' social or economic class. In effect, ESEA was developed under the principle of redress and equity, which established that children from low-income homes required more educational services and resources than children from more affluent homes, and, thus, their school districts and schools would receive more federal funding for primary and secondary school education than their more affluent counterparts. In 1965 when ESEA became law, there was a large achievement gap explained by race and poverty. Over time, Title I of ESEA focused specifically on improving the academic achievement of children from low-income backgrounds and closing the achievement gap between children from different demographic groups, including those from rural and immigrant families as well as families with limited English proficiency. It also held school districts and schools accountable for closing the achievement gap between test-takers' observed and expected performance.

The ESEA opened the way for new and increased uses of large-scale tests to evaluate education and social programs funded by the federal government. In fact, the use of tests and other forms of evaluation became necessary to document the progress (or lack thereof) that schools were making toward improving the education of children from low-income backgrounds. The arrival of NAEP in 1969 was timely for capturing national-level performance data on students from backgrounds that were the focus of ESEA.

Beginning in the 1970s and beyond, reauthorizations of and amendments to ESEA coupled with landmark court rulings, focused on the educational inequities of students with disabilities (Individuals with Disabilities Education Act of 1974) and English learners (Every Student Succeeds Act [ESSA], 2015, Title III, Part A). Consequently, NAEP was asked to provide more information about student group performances so that government and education officials would have a stronger basis for making judgments about the adequacy of education services and the attainment of education goals for these groups (e.g., attainment of English language proficiency and goals related to state academic standards). NAEP's original design could not accommodate the increasing demands for data about these educationally important populations and issues. NAEP's original reporting design measured change on individual items and tasks but not for large clusters of items that could constitute a construct or content area. Hence, NAEP's design needed to change to meet the data and information demands of government and education leaders as well as the diverse sets of constituent stakeholders.

The first major redesign of NAEP took place in 1984, when the responsibility for its development and administration was moved to ETS. By contracting a testing company to carry out the actual assessment, the technical design and administration of NAEP was and

still is situated at the forefront of large-scale assessment methodology. The redesign of NAEP was characterized by changes in sampling, objective setting, item development, data collection, and analysis (Messick et al., 1983). Tests were administered by age to facilitate the continuation of trend data and by grade groupings to accommodate how schools are organized for educating children. In addition, summary scores were provided for subject or content area linkages to educational policies and standards. This was the first of several redesigns for NAEP (Pellegrino et al., 1999).

Governance for NAEP: The National Assessment Governing Board

The National Assessment Governing Board (NAGB) was formed in 1988 and charged with setting many of NAEP's policies. These policies include defining the content and format of the assessments: the assessment frameworks; setting achievement levels standards and guiding the development of what it means to be basic, proficient, or advanced; and reporting and disseminating the initial release of NAEP results. Board members must be bipartisan, and the board must include multiple stakeholders, such as educational measurement experts, educators, and community members. The membership rules help ensure that NAEP remains independent and reflective of diverse perspectives and goals. To oversee NAEP, the NAGB works with the NCES and NAEP contractors who design and administer the test. NCES implements the policies articulated by the NAGB and is responsible for the full production and administration of NAEP as well as contractual relationships with the NAEP Alliance and other contractors. NCES also reviews and releases all technical reports generated by members of the NAEP Alliance (Buckendahl et al., 2009). This organizational structure, as defined by the NAGB, has important political and practical benefits for NAEP.

Evaluating the Validity and Utility of NAEP: The NAEP Validity Studies Panel

The NAEP Validity Studies (NVS) Panel was formed by the American Institutes for Research® under contract with the NCES in 1995 to provide a technical review of NAEP plans and products and identify technical concerns and promising techniques worthy of further study and research. The NVS Panel serves as an independent voice and collaborates appropriately with the NAGB, the NCES, and the NAEP Alliance (of contractors) to consider issues related to the validity and utility of NAEP. During the early years of its existence, the NVS Panel published several studies on aspects of NAEP development and implementation (Bock & Zimowski, 1998; Chromy, 1998; Durán, 2000; Hedges & Vevea, 1997; Jaeger, 1998; Mullis, 1997; Pearson & Garavaglia, 1997). Then, in 2003, the NVS Panel addressed NAEP's need for a comprehensive agenda for validity research with the publication of *An Agenda for NAEP Validity Research* (Stancavage et al., 2003).

The validity research agenda developed by the NVS Panel in 2003, included a broad framework of research priorities consisting of six categories: (a) the constructs measured within each of NAEP's subject domains; (b) the manner in which these constructs are measured; (c) the representation of the population; (d) the analysis of data; (e) the reporting and use of NAEP results; and (f) the assessment of trends. Although these categories were presented as part of a validity research agenda, a close examination reveals the presence of equity concerns, though they are not explicitly stated.

The purpose of the validity research agenda was to prepare a systematic analysis of the domain of validity threats and identify the most urgent validity research priorities for NAEP. The importance of establishing the validity research agenda was obvious for at least two reasons, especially with the passage of Public Law 107-110 (No Child Left Behind) in 2002. First, NAEP was expected to have a greater role in helping states judge the adequacy of their yearly progress for their overall student population, as well as for important subsets of student groups defined demographically in the intended testing population. Second, potential harm could result from an education policy perspective if the research studies were not done, and it was later discovered that the hypothetical threats to the validity of NAEP were real. That was over 20 years ago.

IMPROVING EQUITABLE MEASUREMENT AND REPORTING IN NAEP

We are now in the third decade of the 21st century, so we can reflect on the lessons learned from validity research conceived and conducted during the first two decades of this century. It is evident that the playbook for improving equitable measurement and reporting in NAEP is evolving. Demographic changes in the United States in the past 20 years with respect to racial/ethnic and cultural diversity are evident. These changes have implications for how NAEP is designed, developed, administered, scored, and reported. It is time, once again, for the NVS Panel to systematically analyze the validity threats to NAEP but with an explicit reference to concerns about equity and how equitable features can be implemented appropriately throughout the assessment process from assessment design and administration to analysis and reporting. It is time for the panel to produce "An Agenda for NAEP Validity Research: Part II."

The following are some suggestions about (a) how equitable features can be implemented in NAEP, (b) some possible research studies, and (c) who or what directly or indirectly benefits from improving equity in NAEP.

Random Sample for Representativeness

Ensure that (a) the random sample selected to participate in NAEP is "representative" of the nation's school population in important ways and (b) any comparisons of the performances of subgroup samples satisfy the requirements of comparability.

NAEP is designed to report results at the national and state levels and for selected urban districts without requiring every student in every school to take the assessment. Part A of the NAEP Law (Public Law 107-279, Section 303) requires that NAEP "*use a random sampling process* which is consistent with relevant, widely accepted professional assessment standards and *that produces data that are representative on a national and regional basis*" [emphasis added]. NAEP has a very robust plan for selecting representative samples of students to participate in the assessment. The process begins with the selection of schools, both private and public, which considers the location, size, and racial/ethnic composition of the school. Once a school has been selected for either a state or national assessment, students within the school are selected for each subject area based on grade level (4, 8, or 12). From this list of selected elementary, middle, and high schools, a sample of students is randomly selected to participate in the assessment. Every student in the sampled school is eligible to be selected. Then, individual students are assigned to a single subject area to answer questions on a subtest of NAEP items. A technique called balanced incomplete block spiraling or matrix sampling is used to determine which groups of different items are systematically arranged in test booklets to ensure that the entire content domain for the subject area is covered, but each individual student completes only a fraction of the items.

Even though the sampling plan is rigorous and intentional at all phases of the process, sampling error still exists. In terms of checks and balances, two important questions are posed.

1. *Is the ability distribution for a subgroup being assessed in a particular subject area (e.g., reading, mathematics, science) representative of the ability distribution of that subgroup in that content area in the general population?* This question does not compare the ability distributions of

different demographic subgroups. Rather, this question compares the ability distribution of a randomly selected demographic subgroup for a NAEP content area with that group's own ability distribution in that content area in the general population or in a particular state or urban district. The question is asked mainly for the Main NAEP because Main NAEP results are reported as *The Nation's Report Card*; however, the same concern applies, respectively, to interpretations of results based on samples drawn for the state NAEP and the Trial Urban District Assessments (TUDAs).

2. *Are the assumptions of construct invariance or measurement invariance supported for making comparisons and drawing valid conclusions or inferences about the test results of diverse subgroups defined by race/ethnicity, socioeconomic status, or culture?* This question is about the legitimacy of comparing the test results of selected diverse demographic subgroups for a NAEP content area when important assumptions for doing have not been established.

Valid comparisons of diverse subgroups require assurance of construct or measurement comparability or invariance. Otherwise, "apples" are compared with "oranges," such that comparative conclusions are suspect. Lack of construct or measurement invariance is a validity issue. Boer et al. (2018) noted that evidence of measurement comparability for cross-cultural comparisons has been promoted and advocated for many years; however, there are very few references to it in the research literature (e.g., Altarriba, 1993; Guenole & Brown, 2014; Steenkamp & Baumgartner, 1998; van de Vijver & Leung, 1997; Vanderberg & Lance, 2000). Psychometric tools can help establish measurement comparability and aid in drawing valid inferences about cross-cultural differences and rule out alternative explanations related to bias. However, the use of these tools and techniques has received little attention or application despite the increased interest in and prevalence of cross-cultural research.

Research Questions

1. For a given subgroup, does the ability distribution of students sampled for NAEP match the ability distribution for that subgroup in that subject area in the general population?
2. Does NAEP provide credible evidence of construct or measurement comparability (construct or measurement invariance) for subgroups that are racially/ethnically, socially, or culturally diverse? If so, on what bases (statistical and otherwise) are the subgroups deemed to be comparable?

Equity Suggestion

After NAEP is administered, compare what is known about the ability distribution of a demographic subgroup randomly selected to participate in NAEP with what is known about the ability distribution of that group overall. These within-subgroup comparisons and the supporting evidence are conducted to ensure that between-subgroup comparisons of performance on NAEP are fair and meaningful.

Equitable Benefits

States often are ranked based on their students' performance on NAEP. Individual states will be able to compare the distributions of ability on the construct(s) being measured for each subgroup in their states with those of samples of the subgroups in the testing population for NAEP. Demographic diversity differs among states. Satisfying answers to

these research questions address equity issues among states and will help states and their stakeholders interpret the meaning and consequences of the publicly reported rankings.

Suggested Readings (by publication date)

- Linn et al. (2004)
- McLaughlin et al. (2005a, 2005b)
- Grissmer (2007)

Reflecting Current Standards and Practices

Ensure that the frameworks for content areas for the Main NAEP reflect current content standards and practices.

The NAEP framework is the blueprint that guides the development of the assessment and the content to be assessed. NAEP does not exist in a vacuum; its purpose is to provide, in a timely manner, a fair and accurate measurement of student academic achievement and reporting of trends in several subject areas (see Section 303 of the NAEP Law). NAEP also is expected to reflect the lofty goals of standards-based education currently embraced by states. In doing so, NAEP must determine and measure what students know and can do, as well as what students should know and be able to do as productive citizens of the United States and the world. The NVS Panel has published white papers about the adequacy of the NAEP frameworks as well as their alignment with states' standards (e.g., Common Core State Standards) and assessments in the subject areas of mathematics (Daro et al., 2007; Hughes et al., 2013), English language arts (Wixson et al., 2013), science (Pellegrino, 2021a), and civics and U. S. History (O'Malley & Norton (2022) and in monitoring trends (Shepard, 2022).

Ultimately, student achievement results from what is taught and learned in American classrooms. What is taught is based on state or district curricula that contain well-defined learning objectives. What is learned is based on the instructional practices of teachers and whether those practices are sensitive or responsive to the diverse sociocultural backgrounds from which students come, students' opportunities to learn subject area content, and the level of engagement of students in the teaching/learning process. NAEP subject area frameworks, on the other hand, are assessment frameworks, not curriculum frameworks. Because NAEP must fairly and accurately assess students from across the United States, it must represent the union of different curricula; that is, what could be taught in America's classrooms (Shepard et al., 2020). States develop their own state assessments for different purposes—including accountability purposes. These state assessments often come with high-stakes accountability provisions that influence what is taught in the classroom. Therefore, there should be periodic reviews of the extent to which NAEP subject area framework objectives and assessments align with the respective subject area content standards and high-stakes accountability assessments of states and districts. Differences in alignment between states' standards and NAEP frameworks should be noted in statements when comparisons are made about overall group or subgroup performances between states and NAEP or between states.

Research Questions

1. To what extent do newly developed or updated NAEP content frameworks align with current state and district content and practice standards?
2. To what extent do newly developed or updated NAEP content frameworks align with current state and district content assessments?
3. To what extent do NAEP content assessments align with current state and district content assessments?

Equity Suggestion

Conduct periodic reviews of the extent to which NAEP subject area content framework objectives and assessments align with the respective subject area content and learning objectives and high-stakes accountability assessments of states and districts. Examine the extent to which there are differences in emphases in important subdomains (e.g., content, practices) on state versus the Main NAEP assessments.

Equitable Benefits

NAEP and state/district assessments serve different purposes. NAEP serves as a monitoring purpose at the national level; state assessments serve mostly accountability purposes (No Child Left Behind Act of 2002; Every Student Succeeds Act of 2015) at a more local level. Given the national attention that NAEP results receive, states and participating districts (e.g., TUDAs) may be interested in how well NAEP assessments reflect states' content and practice standards and assessments, and, relatedly, how well Main NAEP results relate to what is taught and tested in schools. The results of periodic alignment studies facilitate valid interpretations and uses of NAEP results and protect states, districts, all groups of the test-taking population (especially marginalized subgroups), policymakers, and the public from being exposed to or the subjects of disinformation.

Suggested Readings (by publication date)

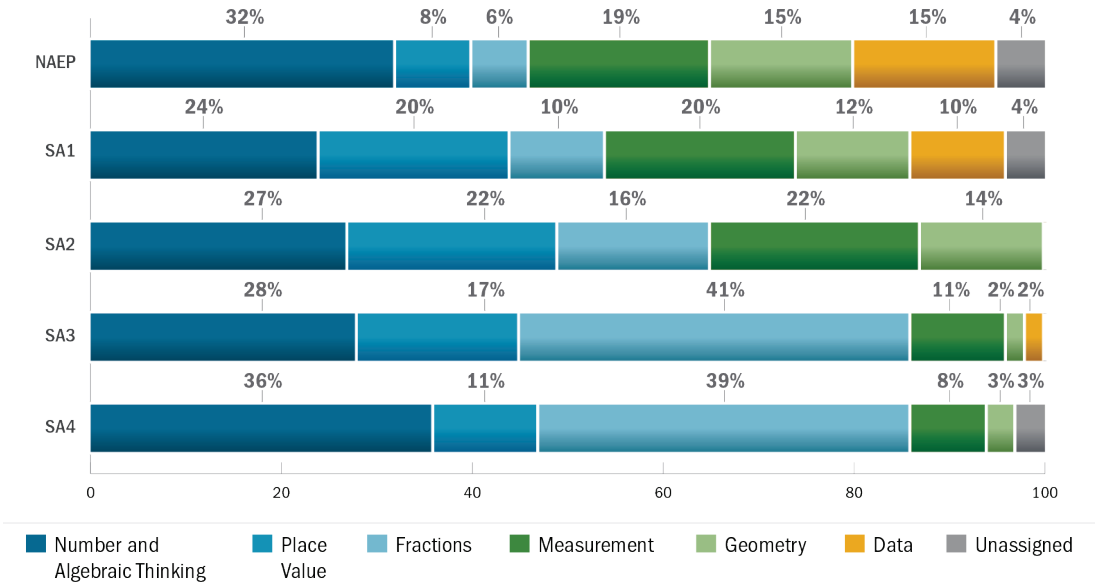
- Daro et al. (2007)
- Stancavage et al. (2009)
- Nellhaus et al. (2009)
- Stancavage and Bohrnstedt (2013)
- Shepard et al. (2013)
- Thissen and Norton (2013)
- Behuniak (2015)
- Daro et al. (2015)
- Valencia et al. (2017)
- Hughes et al. (2019)
- Valencia et al. (2020)

Secondary Analyses

Conduct secondary analyses of NAEP and state assessment data to examine the effects of NAEP framework dimensions (i.e., content emphases, cognitive demand) on assessment results for NAEP, states, and TUDAs.

The educational policy statements of the most recent reauthorizations of the *Elementary and Secondary Act of 1965,* the *No Child Left Behind Act of 2001* and its successor, the *Every Student Succeeds Act of 2015,* implicitly assumed that NAEP content assessments could serve as a common yardstick to monitor the different states' content assessments, which primarily serve accountability purposes. NAEP content assessments are based on the NAEP content frameworks, and states' content assessments are based on states' content and practice standards. Daro et al. (in press) compared the content distribution (e.g., the percentage of total score points for six content domains) on the 2017 NAEP Mathematics assessment for Grade 4 with that found on selected state mathematics assessments. In one representation of the data, 6% of the total score points on NAEP assessment were assigned to fractions, whereas 10%, 16%, 41%, and 39% of the total score points were assigned to fractions, respectively, on four state assessments (see Exhibit 1). Exhibit 1 also shows that 15% of the total score points on NAEP were assigned to the content domain data, whereas 10%, 0%, 2%, and 0% of the total score points were assigned to data, respectively, on four state assessments. The differences in how the two content domains, fractions, and data, were weighted on NAEP versus the state assessments are a good indicator of what was emphasized during classroom instruction in those states. Thus, it could be reasonably concluded that the 2017 NAEP Mathematics Grade 4 assessment tested students on content that many students did not have an opportunity to learn in the content domain of data and underestimated students' knowledge in the content domain of fractions.

Exhibit 1. Content Distribution of the 2017 Grade 4 NAEP and Four Selected State Mathematics Assessments According to the Daro et al. (in press) Classification Scheme



The underestimation of students' knowledge of fractions on NAEP, for example, or any other important content emphases on state accountability assessments, was likely a source of concern among state education agencies and school districts. Superintendents in several TUDAs raised concerns about decreases in students' performances on NAEP between 2003 and 2017. For many states/districts, these years coincided with the adoption of college and career readiness standards. Dogan (2019) conducted a study to determine whether the 2017 NAEP Mathematics Grade 4 mean scores for TUDAs would change if NAEP mathematics subdomains (e.g., numbers, measurement, geometry, data, and algebra) were reweighted to correspond to the content distributions of the respective state assessments for the TUDAs. The results showed that NAEP was underweighted in the subdomain of numbers and overweighted in the subdomain of data (Dogan, 2019; Shepard, 2022). When the appropriate weights were applied to NAEP mean scores for 2013, 2015, and 2017, mean scores increased in nine TUDAs in the Daro et al. (in press) study. Performance gaps also decreased between Black and White students in those TUDAs and states. The increases in NAEP mean scores in the TUDAs and the decreases in the performance gaps between Black and White students were a by-product of equitable measurement that—in essence—provided more accurate estimates of students' knowledge and skills.

Research Questions

1. To what extent are there similarities or differences in content emphases (or cognitive demand) between NAEP and state assessments across time?
2. Do the similarities and differences in content emphases revealed in the Dogan (2019) reanalysis of the NAEP mathematics scores change as test-taker demographics change across time?

Equity Suggestions

Use weights in conducting a reanalysis when concern exists that the content emphases in subdomains on NAEP differ from those on state assessments. Document instances where there is a "real and educationally significant mismatch." If there is a mismatch in the content emphases, monitor how the mismatch changes as test-taker demographics change in the TUDAs. Monitor major changes in content and curricula emphases in the educational system and update NAEP frameworks accordingly and as needed. Gordon (1995) reminds us that "equity speaks to and references fairness and social justice; it requires that the distribution of social resources be sufficient to the condition that is being treated" (p. 363).

Equitable Benefits

This reanalysis, with all its limitations, demonstrates the power of an analytical methodology for examining low-stakes NAEP data from the perspective of what schools and districts value and emphasize in their high-stakes accountability educational systems and assessments.

Suggested Readings (by publication data)

- Dogan (2019)
- Hughes et al. (2019)
- Shepard (2022)

Assessment Access

Design, develop, and administer NAEP in ways that ensure all subgroups of test takers have access to the assessment.

As mentioned earlier, Sireci (2021) noted that a goal of fair and equitable assessment is to level the playing field for all test takers by reconceptualizing standardization in ways that build in flexibility, rather than rigidity, into the testing process. One way to level the playing field for all test takers is to "standardize" on the characteristics of test takers rather than features of the test. Simply put, focus on the characteristics of test takers to (a) ensure they have full access to the test during administration of the assessment, (b) enhance their ability to maximally show what that know and can do, and (c) reduce feelings of anxiety and helplessness. For example, test developers should develop ways to standardize on test takers' levels of engagement, interests, familiarity with devices used to take the assessment, or usability with features of the test or testing context. Because the main purpose of NAEP is to assess and report to the nation what students know and can do, accessibility to the assessment becomes more realistic for all test takers when the testing experience is more student-centered.

Accessibility to NAEP is made possible through accommodations and universal design principles. Many of the traditional accommodations for students with disabilities and English learners are already used in NAEP. Some of these accommodations include extra time, one-on-one administration, directions read aloud only in English, breaks during the test administration, magnification, large print, cueing to stay on task, responds orally to a scribe, presentation and response in Braille, and presentation and response in sign language. Furthermore, NAEP has incorporated universal design elements for all students in digitally based assessments. These elements include, for example, zooming, small group, one-on-one administration, text-to-speech (English) directions only, color contrast, volume adjustment, and closed captioning.

Another area of concern about test takers' accessibility to a test is the use of digitally based forms of test administration. In 2001, NAEP began exploring testing methods and question types that reflect the growing use of technology in education. Since that time, most students have grown up as digital natives. Therefore, in 2017, NAEP transitioned to digitally based assessments in mathematics and reading at Grades 4 and 8. The mode of administration for NAEP was paper since its inception; therefore, the digital transition prompted research on mode effect—paper-based versus digitally based administrations of NAEP. Kitmitto et al. (2018) conducted an evaluation of mode transition in the form of a bridge study. The authors reported no clear evidence of consistent bias in the linked results.

In addition, when it comes to device and interface features of digital test administration, Way and Strain-Seymour (2021) noted that ". . . [a]n equally important potential contributor to device/interface effects is tool familiarity" (p. 37). The researchers reported that usability, familiarity, or the interaction of the two could be a performance-inhibiting experience for a test taker; therefore, the degree to which an assessment can faithfully represent a student's ability may be compromised. Estimates of test takers' familiarity with the device and interface could be used to measure the effects of such on their performance.

In August 2021, the NAGB, which sets policy for NAEP, took another important step toward making NAEP more accessible to all test takers. The NAGB released the *2026*

Reading Framework for the National Assessment of Educational Progress, which describes the content and design of the 2026 Reading Assessment. The policy arm of NAEP acknowledged the influence of test takers' social and cultural experiences on learning and development and, by extension though not explicitly stated, inferred how those experiences influence test takers' interactions with the NAEP reading assessment:

> The 2026 NAEP Reading Framework is updated to reflect three research-based developments that help ensure that the NAEP Reading Assessment remains a useful measure of reading comprehension. *The first is how students' social and cultural experiences shape learning and development, including the learning and development of reading comprehension* [emphasis added]. The second is how reading varies across disciplines. The third regards the use of digital and multimodal texts. (NAGB, 2021, p. 13)

This bold and forward-thinking step by the NAGB provides evidence of its desire to infuse equity in the design and administration of NAEP and interpretations of NAEP results for some of the most marginalized subgroups. Sireci (2021) expressed the point directly in his description of the concept of "understandardization":

> The goal of understandardization is to understand (a) what each student brings to the testing situation in addition to the proficiency measured, (b) how these personal characteristics may interact with testing conditions, and (c) how the testing conditions can be flexible enough to accommodate and account for these potential interactions. (p. 2)

Research Questions

1. In what ways do test takers' sociocultural life experiences affect their different levels of engagement or familiarity with the design of a test?
2. Which testing accommodations or universal design features are most effective in providing access to a test for test takers who have specific physical or cognitive disabilities?

Equity Suggestion

One purpose of NAEP is to measure reading comprehension. One equitable assessment practice for reading comprehension might allow for a variety of reading passages on different topics while controlling for or standardizing on the same complexity and difficulty of all passages for a particular grade level. When students can choose a topic that engages them, standardization becomes student centered. Current standardization practices require that all students read the same passage. Under this latter condition, some students are advantaged because they find the topic of the passage engaging, whereas other students may not be engaged at all. When students are engaged in a reading passage topic, they are more likely to provide better evidence of their comprehension skills. For students who are not engaged in the reading passage, the measurement of their reading comprehension could be underestimated. The underestimates of each student's reading comprehension in the latter group of disengaged students are aggregated across all students in that group and yield an underestimation of the reading comprehension for that group.

Critics of the choice option often cite a seminal research study by Campbell and Donahue (1997): *Students Selecting Stories: The Effects of Choice in Reading Assessment*. The results of that

study indicated that, in general, when students have the option to choose a reading passage rather than assigned one, there is no difference in the reading comprehension scores of the two groups (Campbell & Donahue, 1997). In fact, in cases when students of the same race/ethnicity have a choice versus no-choice option, within race comparisons revealed no significant differences in performance, and in most cases, the no-choice group scored higher than the choice group. There were, however, two exceptions to these results. Black eighth graders and Hispanic 12th graders who had a choice of which passage to read scored higher (not significantly higher) than their peers who did not have a choice. Campbell and Donahue presented these data in tables in the publication of the study but did not address these two exceptions.

Powers and Bennett (1999) conducted a choice versus no-choice study in which examinees who took the Graduate Record Examination were asked to participate in an experimental section of the test for which they would receive no score. Their most consistent finding was that performance was higher, on average, for every item studied when examinees were allowed to choose an item from among a set of items they were required to answer. Powers and Bennett did not, however, conclude that choice is appropriate for every testing context. Instead, they noted as follows: "It seems clear, at least *under some circumstances* [italics mine], examinee choice can have salutary effect on measurement both for the examinee and for the examiner" (p. 276). A search of the choice literature also indicates that when some individuals (especially marginalized groups) have "agency" or the power to choose, it has the potential to motivate them to perform better than perhaps they normally would. When the design of the test influences students' level of engagement with the test, then student engagement becomes a measure of students' accessibility to the test.

Being responsive to the different characteristics that students bring to the testing context will require more resources and research and will add to the cost of developing (research, field testing) and administering (organization of blocks) the assessment. Members of the NAEP Alliance of partners and contractors that assist in developing the NAEP Reading Assessment can begin the process by (a) identifying currently used passages according to topics and organize the topics into clusters/folders and (b) collecting data on what engages students from different subgroups—not just by race and ethnicity (cultural norms) but also where students live (urban, suburban, rural), socioeconomic status (life experiences), and gender (social acculturation) and then use this information to develop items or identify appropriate reading passages.

Equitable Benefits

All subgroups of the test-taking population benefit by reducing measurement error and having more accurate and fairer estimates of subgroup performances defined by demographic variables. Although NAEP provides accommodations for students with disabilities and English learners, by acknowledging how culture influences performance on tests, a larger proportion of the test-taking population in NAEP, especially those identified by race/ethnicity, will have greater access to the test content so that they can maximally demonstrate what they know and can do on NAEP.

Suggested Readings (by publication date)

- Gordon and Shipman (1979)
- Lukhele et al. (1994)
- Hood (1998)
- Weston (2002)
- Chromy and Mosquin (2004)
- Linn et al. (2004)
- Arbuthnot (2011)
- DeStefano and Johnson (2013)
- Bohrnstedt et al. (2018)
- Kitmitto et al. (2018)
- Jewsbury et al. (2020)
- Durán et al. (2020)
- Sireci (2020, 2021)

Reporting Comparisons

Report and highlight all comparisons among racial/ethnic subgroups.

Evaluate the validity of interpretations and inferences about subgroup performances and comparisons made from NAEP data and reports.

Research Questions

1. What are the most salient messages received by the various audiences to which NAEP results are reported?
2. Do the messages received by the various audiences reflect what the Nation's Report Card intends to send?

Equity Suggestion

Highlight the Asian/White gaps in achievement on NAEP along with the much-publicized White/African American and White/Hispanic performance gaps. Asian/Pacific Islander students frequently outperform all other racial/ethnic subgroups on NAEP. However, their superior performances to White test takers are rarely prominently reported. Instead, written reports tend to focus on White/African American and White/Hispanic comparisons, in which the superior performance of Whites compared with these groups are emphasized. Members of the NAEP Alliance as well as researchers should conduct more within-subgroup analyses using information from NAEP surveys to determine which factors may explain why some children within the same racial or ethnic subgroup perform better than others. These factors may differ among racial/ethnic subgroups because of the multiple ways in which subgroups experience life in America.

Equitable Benefits

Some White test takers may be disadvantaged educationally because of inequities. Yet, many of them go unnoticed because they slip under the "assessment radar screen" because of their power and privileged status in America.

Suggested Reading

- https://www.nationsreportcard.gov/

Improving Learning and Learning Environments

Use NAEP results to improve student learning and learning environments for all test takers.

There is no expectation that NAEP results can provide valid and reliable feedback to teachers. That is not NAEP's purpose. There is an expectation that NAEP results will help identify policy implications at the national, state, and district levels that will result in minimizing inequities in education for all American students for whom the assessment is designed.

Research Question

1. How can NAEP's role as a monitor of student achievement at the national level contribute to the improvement of student learning and learning environments for all test takers at the national, state, and district levels in a coherent system of assessment?

Equity Suggestion

There are many different purposes for assessing students (e.g., system monitoring, accountability, selection, placement, certification, diagnosis, classification, performance feedback (for learning), evaluation (program and grading), guidance, research, employment). Identify the role of each purpose, prioritize the importance of each purpose given the mission and goals of the educational system, and allocate appropriate financial resources accordingly and proportionally.

Equitable Benefit

The benefit is a more responsive, coherent, balanced educational system that includes assessments from which all students benefit.

Suggested Reading

- The NAEP Law (2002)
- Stiggins (2004)
- Cizek (2020)
- Gordon (2020)

SUMMARY

Efforts to improve fairness and equity in the design and administration of NAEP as well as the validity and utility of the results have occurred throughout NAEP's history and evolution. There is a process by which equity-minded assessment professionals can successfully infuse equity throughout the assessment process with a view toward informing or educating all stakeholders using transparent methods and practices. These transparent methods and practices include but are not limited to the following:

- Describe and be clear about the current practices used in measurement or reporting and describe the approach thought to be more equitable.
- Explain why that approach or process has the potential to contribute to a fairer assessment experience for all students or how it will minimize measurement error using promising results from well-planned and well-executed research studies.
- Explain how the equitable approach or process described will benefit all stakeholders but not disadvantage anyone (e.g., acknowledge the merits of standardized measurement while also identifying any limitations thereof, given the purpose for measuring educational achievement for all students).
- Identify possible limitations, cautions, or concerns, such as added costs associated with the suggested equitable ways of doing assessment and measurement in NAEP and offer solutions, when possible.
- Conduct related research and pilot studies to evaluate the efficacy of the approach and adjust, as necessary.

Exhibit 2 summarizes the places in the assessment process where it may be possible to implement equity in assessment design, administration, scoring, analysis of data, reporting, and the use of assessment results in NAEP.

Exhibit 2. Equity in NAEP

Assessment process	Equity in measurement and reporting
Assessment design and development	
Representativeness and comparability of NAEP random samples within and across subgroups	Compare the ability distribution of a particular NAEP subgroup sample with the ability distribution of that subgroup in the subject area being assessed in the nation, as well as in the states or districts of interest. States, districts, and subgroups benefit.
Alignment of assessment content with frameworks and content standards	Enable a more "evolutionary" approach to framework updates to ensure construct representation in NAEP.
Development of item pool	Ensure that NAEP items align with current content standards and practices. Allow flexibility in the application of standardization features of the assessment.
Administration of assessment	
Accommodations for students and allow flexibility and reasonableness	Acknowledge and respect diversity in the testing sample and administer the assessment accordingly with appropriate accommodations.
Analysis, reporting, and use of results	
Analysis of data	Provide credible evidence of construct and measurement invariance when comparing subgroup performances based on race/ethnicity/gender, socioeconomic status, disability status, and English language proficiency. Conduct more within-group comparisons to document variability and identify factors that may explain successful performance. No subgroup is a monolith that lacks diversity.
Reporting to multiple audiences	Identify the various audiences to which NAEP results are reported and the rationale for doing so (e.g., policymakers, national and state legislative bodies, chiefs of states, the media, the public at large, school districts, school leaders, teachers, students, parents, professional content organizations). Where appropriate, highlight all differences in performance between subgroups by race/ethnicity, not just the White/Black or White/Hispanic performance gaps.
Use of NAEP results	Identify policy implications that may result in minimizing inequities in education.

Toward Equitable Assessment and Measurement in NAEP: Professional Activism

Measurement experts, test developers, policymakers, and a diverse group of users of assessment results in the testing and measurement communities have an opportunity and an obligation to promote fair and equitable testing practices. It is called "professional self-regulation." Many of the principles that support valid, fair, and equitable assessment are embodied in standards' documents developed by the American Educational Research Association, the National Council on Measurement in Education, and the American Psychological Association (Sireci, 2021). If history is our teacher, it will take a village of bold and determined assessment professionals, including policymakers, test developers, and users to improve equitable measurement and reporting in NAEP. To that end, in March 2022, Peggy G. Carr, commissioner of the NCES, and Lesley Muldoon, executive director of the NAGB, issued a joint statement in which they addressed current and future educational needs exposed because of challenges faced during the COVID-19 pandemic: "For us to move forward, we need to understand . . . how our education system—including assessments—can adapt, innovate, and produce improved student outcomes" (Carr & Muldoon, 2022).

Policymakers

Members of the assessment community who set educational assessment policy for large-scale assessment systems (e.g., NAEP) are in enormously powerful positions. They are a diverse group of American citizens who enact the laws, make the rules, and set the course for how the assessment will operate. The NABG sets policies for NAEP, and in the recent release of the 2026 Reading Framework, which guides the design, development, and administration of NAEP as well as the analysis, interpretation, and use of NAEP results, it was acknowledged:

> The 2026 NAEP Reading Framework is updated to reflect three research-based developments that help ensure that the NAEP Reading Assessment remains a useful measure of reading comprehension. *The first is how students' social and cultural experiences shape learning and development, including the learning and development of reading comprehension* [emphasis added]. The second is how reading varies across disciplines. The third regards the use of digital and multimodal texts." (NAGB, 2021, p. 13).

It remains to be seen how the acknowledgement that students' cultural experiences shape learning, and the development of reading comprehension will manifest itself in the design, development, administration, scoring, analysis, and reporting of results for future NAEP reading assessments.

Sometimes in developing policies, policymakers do not always acknowledge the effect that implementing such policies may have on the lives of individuals the policies are intended to help or the efficiencies they hope to reap during the assessment processes. Sometimes the effect is detrimental. Policymakers should develop a listening ear that carefully considers the preponderance of evidence from high-quality research that reveals the damaging effects of the policies they enact. Furthermore, they must be prepared to apply the principle, "All things are lawful, but not all things are advantageous. All things are lawful, but not all things build up. Let each one keep seeking, not his own advantage, but that of the other person" (*New World Translation of the Holy Scriptures*, 2013, 1 Corinthians 10:23-24). In other words,

policymakers must be willing to amend or change policies, where necessary, that result in more damage than good to educational or assessment systems, as well as the students, teachers, and other stakeholders therein. If policy leaders are supportive of equity-based approaches, there will be evidence of that support in policy statements and actions.

Measurement Specialists and Test Developers

Members of the assessment community who are measurement specialists and test developers include a wide variety of individuals who have training in developmental psychology, psychometrics, educational psychology, cognitive psychology, mathematical sciences, linguistics, curriculum and instruction, sociology, and a variety of content areas. These members of the community are in the "conference rooms" and on the "assembly line" for developing NAEP and other assessments. Their professional activism, and especially that of professionals at NCES and the NAEP Alliance, may be manifested in many ways, including but not limited to whether (a) the research in which they engage includes questions about the efficacy and appropriateness of equity-based approaches to assessment and measurement and the effects of such on the validity of assessment results for all test takers or (b) they "think outside the box" about concepts and constructs long taken for granted but challenged by notions of equity that are a part of their professional culture—topics such as the operationalization of standardization and comparability in a testing context. Measurement specialists and test developers must also be prepared, when appropriate, to look beyond the "numbers," psychometrically speaking, and follow-up on assessment results that may not be statistically significant but may open up new avenues for exploratory research. In addition, university professors and professionals who work in testing companies can serve as equity-centered mentors to junior assessment and measurement professionals who are deciding on a portfolio of research for the present and for years to come.

Users of Assessment Results

Users of assessment results are perhaps the most diverse group in the assessment community. They include all the aforementioned individuals—policymakers, measurement specialists, and test developers—plus teachers, students, parents/guardians, principals, district superintendents, curriculum developers, professional developers, business leaders, and more. Users of assessment results are in a position to do either good or harm. They have an obligation to understand what conclusions can be validly drawn from the data.

NCES can play a pivotal role in educating users for NAEP results from all walks of life about what is in the NAEP portfolio of assessments—the subject areas covered, the design of the assessments, how to interpret achievement levels, whether students' performance on state assessments predict their performance on NAEP, do the main and LTT assessments measure the same thing, and so on. As activists for equitable measurement and reporting in NAEP, users of NAEP results must be sure that their interpretations of the results are accurate and their uses of the results are appropriate, given the intended purpose of the assessment. Their motto should be "do no harm!"

CONCLUSION

The year 2023 marks the 20th anniversary of the release of *An Agenda for NAEP Validity Research* (Stancavage et al., 2003). It is time, once again, to develop a validity research agenda that features equity themes explicitly. Today, more than ever before, the centrality of fairness and equity at all stages of the assessment process, particularly in interpreting and using assessment results in making educational policy decisions and in inferring educational effectiveness, is ever present. An educational system that aims to develop all its human capital is one in which all students have an opportunity to close achievement gaps, not only between each other but also, most importantly, between what they are expected to learn, what they actually learn, and what they are able to demonstrate that they have learned—in a fair and equitable assessment system, of which NAEP plays a major role.

At the beginning of this paper, I posed a question directed at NCES, the NAEP Alliance, and the NAGB in particular, but one that should be seriously considered by all measurement, testing, and assessment professionals:

> Do current practices and methodologies used by NAEP (e.g., design, development, administration, scoring) mostly assume that American students are part of a "melting pot" socially and culturally or do they acknowledge and reflect that America has become more socially and culturally diverse since the inception of NAEP?

Throughout this paper, I attempted to provide credible references to published research studies, the *Standards*, and logical reasoning that it is imperative that we pay attention to America's progress toward a multicultural society and act accordingly, responsibly, and expeditiously within a justice-oriented validity framework (Randall et al., 2022).

As mentioned earlier, Bennett (2023) proffers five provisional principles upon which a working definition and initial theory of equity-based, socioculturally responsive assessment (SCRA) are based. Socioculturally responsive assessment (1) includes problems that connect to the cultural identity, background, and lived experiences of all individuals; (2) allows forms of expression and representation in problem presentation and solution that help individuals show what they know and can do; (3) promotes deeper learning by design; (4) adapts to personal characteristics including cultural identity; and (5) characterizes performance as an interaction among extrinsic and intrinsic factors. The development of a network of empirically testable propositions can be used to better understand how measures designed from the SCRA principles are supposed to operate, for which purposes and in what contexts, what they are expected to achieve, and if they work. Furthermore, achieving widespread operational use of socioculturally responsive assessment will require considerable time, thought, iteration, and skill.

That said, I will not reiterate all the individual recommendations that speak to proposed ways of improving equitable measurement and reporting in NAEP and improving many other steps in the assessment process. Instead, I will conclude with the following three recommendations for NAEP and the field at large.

- *Acknowledge the influence of sociocultural factors on test performance and conduct additional research to estimate the sizes of the effects of different sociocultural factors on the test performance of racially and culturally diverse groups of test takers* (Bennett, 2023; Mislevy, 2018; Basterra et

al., 2011; Randall et al., 2022). The NAGB has moved in this direction by approving the new research-based 2026 Reading Framework, which reflects how students' social and cultural experiences shape learning and development, including learning and the development of reading comprehension. The new reading framework will serve as the basis for developing future NAEP reading assessments. A new validity research agenda will no doubt include studies that will examine the extent to which the goals of the 2026 Reading Framework are reflected in the design of the assessment, item development, the administration of the assessment, the scoring, analyses, and the reporting and use of the results.

- *Revisit and re-educate members in the field, where necessary, about fundamental concepts and terms in testing, assessment, and measurement that have been taken for granted and viewed as the "norm" for decades.* Experts in the field have been acculturated by a "melting pot" society or caste system vision of America (Wilkerson, 2020). In either vision of America, any deviations from the dominant group's cultural norms often are described using negative language (e.g., construct-irrelevant, threat to validity). A new validity research agenda may reveal better ways to communicate about our diversity and capitalize on the opportunity to revise some of the professional language that is used.
- *Re-fashion assessment generally, and standardized testing specifically, to reflect (a) the different purposes for assessment and (b) the multicultural society America is becoming.* The re-fashioning of assessment will provide an opportunity to re-center testing on test takers. This undertaking, if done properly, will likely be complex, costly, and time-consuming, but it is imperative if the field in general, and NAEP in particular, are to remain relevant and useful (Bennett, 2022).

I remain cautiously optimistic but optimistic, nonetheless. I say "cautiously" because I am somewhat aware of the challenges—political and otherwise—that lie ahead. I also am fully aware that discussions about the influence of sociocultural factors on test takers' performance are not new. For the past half century, our acknowledgments of these influences have been more sinusoidal than exponential; that is, we keep coming back to the same discussions. I cite two experiences I had while writing this paper.

- Earlier, I described some of the former days when NAEP was conceptualized from 1963 to 1969.

  > Although both Keppel and Tyler had compatible interest in creating a national assessment to evaluate student learning, they had somewhat different visions about how it should be conceived. According to Jones and Olkin (2004), in *The Nation's Report Card: Evolution and Perspectives*, Tyler believed (a) "that commonly used standardized achievement tests did not provide a valid measure of what children have learned but were designed to rank students"; (b) "the purpose of standardized tests was to identify individual differences in achievement*, not to measure individuals' learning*" [emphasis added]; and (c) "the manner in which standardized tests were scored and reported was not a meaningful way to score and report the achievement of a community" (p. 26). Keppel, on the other hand, wanted to have national data that would meet the intent of the legislation that created the U.S. Department of Education."

Given the way that NAEP currently operates, it is apparent that Tyler's original vision for NAEP did not prevail.

- I came across a special edition of the *Journal of Negro Education* published in summer 1968, the year I graduated from high school. The theme of the issue was "Race and Equality in American Education." I was drawn to "Section 2: The Assessment of Negro Capacity and Achievement" and to an article written by Winton H. Manning, executive director of research and development at the College Entrance Examination Board. The title of the article is "The Measurement of Intellectual Capacity and Performance." As a summary of the article, Manning (1968) wrote the following:

  > The cultural role of language probably underlies the pervasiveness of verbal aptitude in present methods of assessing developed abilities, but *a new direction for testing requires fundamental changes in the conception of measurement arising from new students of language and cognitive development in children* [italics mine] . . . The circle of our psychometric understanding must be enlarged by what is happening outside that circle, especially in studies of psycholinguistics, thinking, and related processes. (p. 267)

As I have said in many ways before: "Been there, done that!" So why be optimistic now? I can be optimistic because we are at an inflection point in American society, in which discussions and activity about diversity, equity, inclusion, fairness, and justice abound. Therefore, it is encouraging to note that in the publicly released joint statement issued by two leaders who are responsible for the future of the NAEP program, Carr and Muldoon pledged to work together and with their stakeholders "to strengthen and re-imagine NAEP, building on its reputation for rigor, quality, and independent, scientific integrity" (Carr & Muldoon, 2022). The two leaders will collaborate on a research and development agenda that will be guided by the six principles of utility, relevance, adaptability, equity, efficiency, and quality. Each principle was defined in the joint statement; however, for my purposes, I prioritized them in the following way:

- **Quality.** Innovations that push the new frontier also must uphold the technical rigor, integrity, and validity of NAEP.
- **Adaptability/Equity.** These two principles are equally important. A belief in and commitment to a culture of adaptability is a prerequisite to a commitment to acting in equitable ways.
  - **Adaptability.** It is important that NAEP—which, by design, makes changes carefully and methodologically—become more adaptable and nimbler to ensure its ongoing **relevance**. This includes more flexible administration and updating content, as needed.
  - **Equity.** NAEP must lead the field in identifying new, empirically based methods to advance equity in assessment. The United States is a nation of incredible diversity, and assessment data must capture, to the extent possible, the experiences of students of every race, gender, culture, and ethnicity - as well as students with disabilities and English learners - looking deeper at what may explain widening gaps and insights into how to close them. With diligent effort and commitment, we can get closer to understanding the "whys" of inequities without crossing the line of unjustifiable cause and effect.

- **Utility.** For data to be useful, they must paint a deeper picture of the educational experiences of all students and meet with policymakers and other stakeholders in states and districts to hear what they need most to use the results to support meaningful improvements for all students while also reducing burden on schools, reducing costs, seeking opportunities to create new efficiencies, and maintaining NAEP as the "gold standard."

The re-fashioning of NAEP by applying these principles under the leadership of NCES and NAGB will yield an outstanding contribution to American education for all groups of people in ways that are immeasurable.

REFERENCES

Abedi, J., & Ewers, N. (2013). *Accommodations for English learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

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

Arbuthnot, K. (2011). *Filling in the blanks: Understanding standardized testing and the Black/White achievement gap*. Information Age Publishing.

Arbuthnot, K. (2015, October 3). *The four tiers of fairness: Examining the complexities of test fairness and the assessment of diverse populations of test takers* [Paper presentation]. Annual Conference of the National Association of Multicultural Education, New Orleans, LA, United States.

Arbuthnot, K. (2020). Reimagining assessments in the post pandemic era: Creating a blueprint for the future. *Educational Measurement: Issues and Practice*, *39*(3), 97–99. https://doi.org/10.1111/emip.12376

Altarriba, J. (Ed.). (1993). *Cognition and culture: A cross-cultural approach to cognitive psychology*. Elsevier Science Publishers B.V.

Barnes, E. (1972). *I.Q. testing and minority children: Imperatives for change* [Technical paper]. University of Connecticut, National Leadership Institute Teacher Education/Early Childhood. https://files.eric.ed.gov/fulltext/ED078006.pdf

Basterra, M., Trumbull, E., & Solano-Flores, G. (Eds.). (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. Routledge.

Behuniak, P. (2015). *Maintaining the validity of the National Assessment of Educational Progress in the Common Core based environment*. American Institutes for Research. https://www.air.org/sites/default/files/Validity-NAEP-Common-Core-Environment-March-2015.pdf

Bennett, R. E. (2021). *Rethinking assessment for equitable education* [The Edmund Gordon Centennial Conference]. Teachers College.

Bennett, R. E. (2022). The good side of COVID-19. *Educational Measurement: Issues and Practice*, *41*(1), 1–3. https://doi.org/10.1111/emip.12496

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, 1-22. https://doi.org/10.1080/10627197.2023.2202312

Berman, A., Haertel, E., & Pellegrino, J. (2020). *Comparability issues in large-scale assessment: Issues and recommendations*. National Academy of Education Press. https://naeducation.wpenginepowered.com/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf

Bock, R. D., & Zimowski, M. F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP.* American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/Bock_twostage_0.pdf

Boer, D., Hanke, H., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Bohrnstedt, G., Kitmitto, S., & Park, B. J. (2017). *Initial tables from the 2015 computer access and familiarity study.* American Institutes for Research.

Brown v. Board of Education of Topeka, 347 U.S. 483 (1954). https://www.oyez.org/cases/1940-1955/347us483

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L., & Wells, C. S. (2009). *Evaluation of the National Assessment of Educational Progress: Final report.* U.S. Department of Education. https://www2.ed.gov/rschstat/eval/other/naep/naep-complete.pdf

Campbell, J. R., & Donahue, P. L. (1997). *Students selecting stories: The effects of choice in reading assessment: Results from the "The NAEP Reader" special study of the 1994 National Assessment of Educational Progress* (NCES 97-491). Educational Testing Service. https://nces.ed.gov/nationsreportcard/pdf/main1994/97491.pdf

Carr, P., & Muldoon, L. (2022, March 11). *NAEP: Meeting today's needs and building a national assessment for the future* [Press release]. https://nces.ed.gov/whatsnew/commissioner/remarks2022/3_11_2022.asp

Chromy, J. R. (1998). *The effects of finite sampling on state assessment sample requirements.* American Institutes for Research. https://www.air.org/sites/default/files/2022-07/Effects-Finite-Sampling-Corrections-State-Assessment-Sample-1998-508.pdf

Chromy, J., & Mosquin, P. (2004). *Federal sample sizes for confirmation of state tests in the No Child Left Behind Act.* American Institutes for Research. https://files.eric.ed.gov/fulltext/ED506848.pdf

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and uses*. Routledge.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115–124. https://www.jstor.org/stable/1434406

Cole, N. (1973). Bias in selection. *Journal of Educational Measurement*, *10*(4), 237–255. https://eric.ed.gov/?id=EJ090322

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–220). Macmillan.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, *8*(2), 71–82. https://www.jstor.org/stable/1433960

Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP mathematics assessment: Grades 4 and 8*. American Institutes for Research. https://files.eric.ed.gov/fulltext/ED499213.pdf

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of 2015 NAEP mathematics items at grade 4 and 8 to Common Core State Standards (CCSS) for Mathematics*. American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf

Daro, P., Hughes, G. B., Stancavage, F., Shepard, L., Webb, D., Kitmitto, S., & Tucker-Bradway, N. (in press). *A comparison of the 2017 NAEP mathematics assessment with current-generation state assessments in mathematics: Expert judgment study*. American Institutes for Research.

Davis, A. (1948). *Social class influences upon learning*. Harvard University Press.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, *27*(2), 14–24. https://doi.org/10.2307/1176194

DeStefano, L., & Johnson, J. (2013). *Study of the feasibility of a NAEP mathematics accessible block alternative*. American Institutes of Research. https://www.air.org/sites/default/files/2022-07/Feasibility-NAEP-Mathematics-Accessible-Block-Alternative-August-2013-508.pdf

Diana v. State Board of Education, Civ. Act. No. C-70-37 (N.D. Cal., 1970, further order, 1973)

Dixon-Román, E., & Gordon, E. W. (2012). *Thinking comprehensively about education: Spaces of educative possibility and the implications for public policy*. Routledge.

Dixon-Román, E., & Gergen, K. J. (2013). *Epistemology and measurement: Paradigms and practices: Part I. A critical perspective on the sciences of measurement*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/dixon_roman_gergen_epistemology_measurement_paradigms_practices.pdf

Dogan, E. (2019). Appendix: Analysis of recent NAEP TUDA mathematics results based on alignment to state assessment content. In G. Hughes, P. Behuniak, S. Norton, S. Kitmitto, & J. Buckley (Eds.), *NAEP Validity Studies Panel responses to the reanalysis of TUDA mathematics scores.* American Institutes for Research. https://files.eric.ed.gov/fulltext/ED599642.pdf

Durán, R. P. (2000). *Implications of electronic technology for the NAEP assessment.* American Institutes for Research. https://www.air.org/sites/default/files/2022-04/Implications-of-Electronic-Technology-for-the-NAEP-Assessment-NVS-Panel-2000.pdf

Durán, R. P. (2011). Ensuring valid educational assessments for ELL students: Scores, score interpretation, and assessment uses. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 115–142). Routledge.

Durán, R. P., Zhang, T., Sanosa, D., & Stancavage, F. (2020). *Effects of visual representations and associated interactive features on student performance on the National Assessment of Educational Progress (NAEP) pilot science scenario-based tasks.* American Institutes for Research. https://files.eric.ed.gov/fulltext/ED606244.pdf

ETS. (2014). *Standards for quality and fairness.* https://www.ets.org/pdfs/about/standards-quality-fairness.pdf

Faulkner-Bond, M., & Soland, J. (2020). Comparability when assessing English learner students. In A. Berman, E. Haertel, & J. Pellegrino (Eds.), *Comparability issues in large-scale assessment: Issues and recommendations* (pp. 149–175). National Academy of Education Press. https://naeducation.wpenginepowered.com/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf

Ford, D. Y. (2004). *Intelligence testing and cultural diversity: Concerns, cautions, and considerations.* University of Connecticut, National Research Center on the Gifted and Talented. https://files.eric.ed.gov/fulltext/ED505479.pdf.

Ford, D. Y. (2007). Intelligence testing and cultural diversity: The need for alternative instruments, policies, and procedures. In J. L. VanTassel-Baska (Ed.), *Alternative assessments with gifted and talented students* (pp. 107–128). Prufrock Press and the National Association for Gifted Children.

Gándara, P. (2017). The potential and promise of Latino students. American Educator, 41(1), 4. https://files.eric.ed.gov/fulltext/EJ1137807.pdf

Gandara, F., & Randall, J. (2019). Assessing mathematics proficiency of multilingual students: The case of translanguaging in the Democratic Republic of the Congo. *Comparative Education Review*, *63*(1), 58–78. https://doi.org/10.1086/701065

Gordon Commission. (2013). *To assess, to teach, to learn: A vision for the future of assessment* [Technical report]. Educational Testing Service. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *Journal of Negro Education*, *64*(3), 360–372. https://doi.org/10.2307/2967215

Gordon, E. W. (1999). *Education & justice: A view from the back of the bus*. Teachers College Press.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*(3), 72–78. https://doi.org/10.1111/emip.12370

Gordon, E. W., & Shipman, S. (1979). Human diversity, pedagogy, and educational equity. *American Psychologist*, *34*(10), 1030–1036. https://doi.org/10.1037/0003-066X.34.10.1030

Gould, S. J. (1996). *The mismeasurement of man*. W.W. Norton and Company.

Grissmer, D. (2007). *Estimating effects of non-participation on State NAEP scores using empirical methods*. American Institutes for Research. https://www.air.org/sites/default/files/2022-12/AIRNVSGrissmerrevised.pdf

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, Article 980. https://doi.org/10.3389/fpsyg.2014.00980

Hedges, L. V., & Vevea, J. L. (1997). *A study of equating in NAEP*. American Institutes for Research. https://www.air.org/sites/default/files/A-Study-of-Equating-in-NAEP-NVS-Panel-1997.pdf

Helms, J. E. (1992). Why is there no study of culture equivalence in standardized cognitive ability testing? *American Psychologist*, 47(9), 1083–1101. https://doi.org/10.1037/0003-066X.47.9.1083

Hilliard, A. G. (1979a). *Standardized testing and the African American: Building assessor competence in systematic assessment*. San Francisco State University, School of Education.

Hilliard, A. G. (1979b). Standardization and cultural bias impediments to the scientific study and validation of "intelligence." *Journal of Research and Development*, *12*(2), 47–58. https://eric.ed.gov/?id=EJ205807

Hilliard, A. G. (1991). *Testing African American students*. Aaron Press.

Hirsch, E. D., Jr. (1996). *The schools we need: And why we don't have them*. Random House.

Hobson v. Hansen, 269 F. Supp. 401 (D.D.C., 1967). https://law.justia.com/cases/federal/district-courts/FSupp/269/401/1800940/

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *Journal of Negro Education*, *67*(3), 187–196. https://doi.org/10.2307/2668188

Hughes, G. B., Daro, P., Holtzman, D., & Middleton, K. (2013). A study of the alignment between the NAEP mathematics framework and the Common Core State Standards for Mathematics (CCSS-M). In F. Stancavage & G. Bohrnstedt (Eds.), *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress* (pp. 9–86). American Institutes for Research. https://www.air.org/sites/default/files/2021-06/NAEP_Validity_Studies_combined_report_updated_9-19-13_0.pdf

Hughes, G. B., Behuniak, P., Norton, S., Kitmitto, S., & Buckley, J. (2019). *NAEP Validity Studies Panel responses to the reanalysis of TUDA mathematics scores.* American Institutes for Research. https://files.eric.ed.gov/fulltext/ED599642.pdf

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress.* American Institutes for Research. https://www.air.org/sites/default/files/Reporting-the-Results-of-the-National-Assessment-of-Educational-Progress-NVS-Panel-1998.pdf

Jencks, C., & Phillips, M. (Eds). (1998). *The Black-White test score gap*. Brookings Institution Press.

Jensen, A. R. (1980). *Bias in mental testing*. Free Press.

Jewsbury, P., Finnegan, R., Xi, N., Jai, Y., Rust, K., Burg, S. (with Donahue, P., Mazzeo, J., Cramer, B., & Lin, A.). (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study.* U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Johnson, S. T. (1979). *The measurement mystique: Issues in selection for professional schools and employment.* Howard University, Institute for the Study of Educational Policy.

Johnson, S. T. (1980). Major issues in measurement today: Their implications for Black Americans. *Journal of Negro Education*, *49*(3), 253–262.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives.* Phi Delta Kappa Educational Foundation.

Kitmitto, S., Bohrnstedt, G., Park, B. J., Bertling, J., & Almonte, D. (2018). *Developing new indices to measure digital technology access and familiarity.* American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/Developing-New-Indices-to-Measure-Digital-Technology-Access-and-Familiarity-October-2018.pdf

Larry P. v. Riles, 495 F. Supp. 926 (N.D. California 1979). https://law.justia.com/cases/federal/district-courts/FSupp/495/926/2007878/

Lau v. Nichols, 414 U.S. 563 (1974). https://www.oyez.org/cases/1973/72-6520

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education*, *67*(3), 268–279. https://doi.org/10.2307/2668195

Linn, R. (1973). Fair test use in selection. *Review of Educational Research*, *43*(2), 139–161. https://doi.org/10.3102/00346543043002139

Linn, R., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, *18*(2), 109–118. https://www.jstor.org/stable/1434652

Linn, R., McLaughlin, D., Jiang, T., & Gallagher, L. (2004). *Assigning adaptive NAEP booklets based on state assessment scores: A simulation study of the impact on standard errors.* American Institutes for Research. https://www.air.org/sites/default/files/2022-11/NVS-Assigning-Adaptive-NAEP-Booklets-2004-508.pdf

Lukhele, R., Thissen, D. & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected item on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–280. https://www.jstor.org/stable/1435268

Manning, W. H. (1968). Intellectual capacity and performance. *Journal of Negro Education*, *37*(3), 258–267. Howard University.

McLaughlin, D. H., Scarloss, B. A., Stancavage, F. B., & Blankenship, C. D. (2005a). *Using state assessments to assign booklets to NAEP students to minimize measurement error: An empirical study in four states.* American Institutes for Research. https://www.air.org/sites/default/files/2021-06/McLaughlin_MeasurementError_final_0.pdf

McLaughlin, D. H., Scarloss, B. A., Stancavage, F. B., & Blankenship, C. D. (2005b). *Using state assessment to impute achievement of students absent from NAEP: An empirical study in four states.* American Institutes for Research. https://files.eric.ed.gov/fulltext/ED506846.pdf

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

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

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era.* Educational Testing Service. https://eric.ed.gov/?id=ED236156

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* Routledge. https://doi.org/10.4324/9781315871691

Mullis, I. V. S. (1997). *Optimizing state NAEP: Issues and possible improvements.* American Institutes for Research. https://www.air.org/sites/default/files/Optimizing-State-NAEP-Issues-and-Possible-Improvements-NVS-Panel-1997.pdf

National Academy of Education. (2021). *Educational assessments in the COVID-19 era and beyond*. https://naeducation.org/wp-content/uploads/2021/02/Educational-Assessments-in-the-COVID-19-Era-and-Beyond.pdf

National Assessment Governing Board. (2021). *Reading framework for the 2026 National Assessment of Educational Progress*. U.S. Department of Education. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/reading/2026-reading-framework/naep-2026-reading-framework.pdf

Nellhaus, J., Behuniak, P., & Stancavage, F. (2009). *Guiding principles and suggested studies for determining when the introduction of a new assessment framework necessitates a break in trend in NAEP*. American Institutes for Research. https://www.air.org/sites/default/files/2022-11/NVS-Nelhaus-Principles-Visual-Representations-2020-508.pdf

New World Translation of the Holy Scriptures (2013). 2023 Watch Tower Bible and Tract Society of Pennsylvania, 1 Corinthians 10, 23-24. https://www.jw.org/en/library/bible/nwt/books/1-corinthians/10/#v46010023

Office of Civil Rights. (2021). *Title IX of the Education Amendments of 1972*. U.S. Department of Health and Human Services. https://www.hhs.gov/civil-rights/for-individuals/sex-discrimination/title-ix-education-amendments/index.html

O'Malley, F., & Norton, S. (2022). *Maintaining the validity of the NAEP frameworks and assessments in civics and U.S. history*. American Institutes for Research. https://www.air.org/sites/default/files/2022-04/Maintaining-Validity-of-NAEP-Frameworks-Civics-US-History-508-Jan-2022b.pdf

Pearson, P. D., & Garavaglia, D. R. (1997). *Improving the information value of performance items in large scale assessments*. American Institutes for Research. https://www.air.org/sites/default/files/Improving-the-Information-Value-of-Performance-Items-in-Large-Scale-Assessments-NVS-Panel-1997.pdf

Pellegrino, J. W. (2021a). *NAEP Validity Studies white paper: Revision of the NAEP science framework and assessment*. American Institutes for Research. https://www.air.org/sites/default/files/2021-11/Revision-of-the-NAEP-Science-Framework-and-Assessment-October-2021.pdf

Pellegrino, J. (2021b). *Towards "Next Generation" affirmative and formative assessments for learning, and in the service of learning* (Discussant). The E. W. Gordon Centennial Conference. Teachers College, Columbia University. https://youtu.be/9gSbb0So7MA

Pellegrino, J., Jones, L. R., & Mitchell, K. J. (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. National Research Council. https://eric.ed.gov/?id=ED446096

Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education*, *12*(3), 257–279. https://doi.org/10.1207/S15324818AME1203_3

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. https://doi.org/10.1111/emip.12429

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. https://doi.org/10.1080/10627197.2022.2042682

Royer, J. M., & Carlo, M. S. (1993). Assessing language comprehension skills in cross-cultural settings. In J. Altarriba (Ed.), *Cognition and culture: A cross-cultural approach to cognitive psychology* (pp. 157–175). Elsevier Science Publishers B.V.

Schmidt, F. L., & Hunter, J. E. (1974). Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist*, *29*(1), 1–8. https://eric.ed.gov/?id=EJ093742

Sharif v. New York State Education Department, 709 F. Supp. 345 (S. D. New York, 1989). https://law.justia.com/cases/federal/district-courts/FSupp/709/345/1586898/

Shepard, L. (2022). *White Paper: NAEP framework and trend considerations.* American Institutes for Research. https://www.air.org/sites/default/files/2022-11/NAEP-Framework-and-Trend-Considerations-October-2022rev-508.pdf

Shepard, L., Daro, P., & Stancavage, F. (2013). The relevance of learning progressions for NAEP. In Stancavage & Bohrnstedt (Eds.), *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress* (pp. 135–252). American Institutes for Research. https://www.air.org/sites/default/files/2021-06/NAEP_Validity_Studies_combined_report_updated_9-19-13_0.pdf

Shepard, L. A., Kitmitto, S., Daro, P., Hughes, G., Webb, D. C., Stancavage, F., & Tucker-Bradway, N. (2020). *Validity of the National Assessment of Educational Progress to evaluate cutting-edge curricula.* American Institutes for Research. https://www.air.org/sites/default/files/2022-04/Validity-of-NAEP-to-Evaluate-Cutting-Edge-Curricula-2020.pdf

Sireci, S. G. (2020). Standardization and understandardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, *40*(1), 7–16. https://doi.org/10.1111/emip.12415

Sireci, S. G., Banda, E., & Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds)., *Handbook of accessible instruction and testing practices* (pp. 231–246). Springer.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. del Rosario Basterra, E. Trumbull, and G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3–21). Routledge.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573. https://doi.org/10.1002/tea.1018

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in testing of English-language learners. *Educational Researcher*, *32*(2), 3–13. https://doi.org/10.3102/0013189X032002003

Stancavage, F., Beaton, A. E., Behuniak, P., Bock, R. D., Bohrnstedt, G. W., Champagne, A., Chromy, J. R., Cole, S., DeMauro, G., Duran, R. P., Grissmer, D., Hedges, L., Hughes, G. B., McLaughlin, D. H., Mullis, I. V. S., Pearson, P. D., & Shepard, L. (2003). *An agenda for NAEP validity research*. American Institutes for Research. https://nces.ed.gov/pubs2003/200307.pdf

Stancavage, F., Shepard, L., McLaughlin, D., Holtzman, D., Blankenship, C., & Zhang, Y. (2009). *Sensitivity of NAEP to the effects of reform-based teaching and learning in middle school mathematics*. American Institutes for Research. https://www.air.org/sites/default/files/2022-11/NVS-NAEP-Sensitivity-to-Instruction-2009-508.pdf

Stancavage, F., & Bohrnstedt, G. (Eds.). (2013). *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress*. American Institutes for Research. https://www.air.org/sites/default/files/2021-06/NAEP_Validity_Studies_combined_report_updated_9-19-13_0.pdf

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, *86*(1), 22-27. https://journals.sagepub.com/doi/pdf/10.1177/003172170408600106

Thissen, D., & Norton, S. (2013). What might changes in psychometric approaches to statewide testing mean for NAEP? In F. Stancavage & G. Bohrnstedt (Eds.), *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress* (pp. 253–304). American Institutes for Research. https://www.air.org/sites/default/files/2021-06/NAEP_Validity_Studies_combined_report_updated_9-19-13_0.pdf

Tyler, R. (1966). The development of instruments for assessing educational progress. In *Proceedings of the 1965 Invitational Conference on Testing Problems* (pp. 95–101). Educational Testing Service.

Valencia, S. W., Wixson, K. K., Ackerman, T., & Sanders, E. (2017). *Identifying text-task-reader instructions related to item and block difficulty in the National Assessment of Educational Progress reading assessment.* American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/Identifying-Text-Task-Reader-Interactions-Related-to-Item-and-Block-Difficulty-NAEP-Oct-2017.pdf

Valencia, S. W., Wixson, K. K., Kitmitto, S., & Doorey, N. (2020). *A comparison of NAEP reading and NAEP writing assessments with current-generation state assessments in English language arts: Expert judgement study.* American Institutes for Research. https://www.air.org/sites/default/files/NVS-NAEP-State-Assessment-ELA-Expert-Judgement-Item-Comparison-Report.pdf

van de Vijver, F. J. R., & Leung, K. (1997*). Methods and data analysis of comparative research.* SAGE.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Way, D., & Strain-Seymour, E. (2021). *A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress.* American Institutes for Research. https://www.air.org/sites/default/files/2022-07/Framework-for-Considering-Device-and-Interface-Features-NAEP-March-2021-508.pdf

Weston, T. J. (2002). *The validity of oral accommodation in testing.* American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/weston_finalrevpdf_0.pdf

Wilkerson, I. (2020). *Caste: The origins of our discontent.* Random House.

Williams, R. L. (1970). Danger: Testing and dehumanizing Black children. *Clinical Child Psychology Newsletter*, *9*(1), 5–6.

Williams, R. L. (1972). *The BITCH-100: A cultural-specific test* [Paper presentation]. American Psychological Association Convention, Honolulu, HI, United States.

Wixson, K. K., Valencia, S. W., Murphy, S., & Phillips, G. W. (2013). A study of NAEP reading and writing frameworks and assessments in relation to the Common Core State Standards in English Language Arts. In F. Stancavage & G. Bohrnstedt (Eds.), *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress* (pp. 87–134). American Institutes for Research. https://www.air.org/sites/default/files/2021-06/NAEP_Validity_Studies_combined_report_updated_9-19-13_0.pdf

Zieky, M. (2016). Fairness in test design and development. In N. J. Dorans and L. L. Cook (Eds). *Fairness in educational assessment and measurement* (pp. 9–31). Routledge.

APPENDIX. TOWARD EQUITABLE ASSESSMENT AND MEASUREMENT IN NAEP: LITIGATION

Fairness and equity in assessment cannot be separated from fairness in education (Hirsch, 1996). The birth of the modern civil rights movement was a watershed in American history and marked an important milestone in the history of schooling and the formal preparation of knowledge and skills for jobs and careers. It also had an influence on testing and testing policy and raised new concerns about the design and use of various types of tests and test results in school and the workplace.

Although the 1954 U.S. Supreme Court decision in *Brown v. Board of Education of Topeka* struck down the "separate but equal" doctrine and had a profound effect on schooling and the context in which schooling and education would occur for all children, especially children of color and from diverse linguistic and sociocultural backgrounds, the decision had no immediate or direct consequences for testing. Yet, it set in motion social and ideological forces that would, in the decades to come, bring student testing and the consequences thereof into closer scrutiny and debate and, ultimately, the courts. During this time, a new branch of applied statistics emerged, which had as its aim the analysis of group differences in test scores to determine the potential negative effects of test use in certain kinds of decisions. In fact, concern about the differential performance of racial and ethnic groups, males and females, and individuals from different socioeconomic classes on standardized tests, and the negative impact that improper uses of test scores had on impacted groups, led to a proliferation of litigation and legislation in the 1960s, 1970s, and 1980s.

Hobson v. Hansen (1967) was filed on behalf of a group of Black students in the Washington, D.C., school system, which stemmed from the school system's practice of assigning students to education tracks primarily based on scores from standardized aptitude tests. As a result, a disproportionate number of Black children were assigned to the lower tracks.

Diana v. State Board of Education (1970) was a class action suit on behalf of nine Mexican American children who had been placed in classes for children with intellectual disabilities. Diana, the lead plaintiff, was a Spanish-speaking student in California who was placed in such a class because she scored low on an intelligence test administered to her in English.

Lau v. Nichols (1974) and *Larry P. v. Riles* (1979) extended the findings in *Diana* to Asian American students and African American students, respectively. *Lau* provided a mandate for bilingual education and bilingual education services, whereas *Larry P.* questioned the use of an intelligence test as the sole criterion for determining a student's educational needs (e.g., placement in special education). Like the ruling in *Diana*, the courts in the *Lau* and *Larry P.* cases ruled that fair assessment practices acknowledge students' linguistic and sociocultural backgrounds. In fact, in the latter three cases, when students' cognitive functioning was reevaluated using alternate tests that matched their linguistic or sociocultural background, their performances improved.

In another important "use" case, *Sharif v. New York State Education Department* (NYSED; 1989) raised the important question of whether New York State acted unfairly when it excluded a large proportion of female students from eligibility and the opportunity to receive the prestigious Regents Scholarship and Empire State Scholarships of Excellence because of

a sole reliance on applicants' scores on the SAT. NYSED had experimented with using different criteria for awarding scholarships, including scores on a specially designed test to assess achievement in college preparatory courses, as well as high school grade point averages. However, in 1989, because of budgetary constraints, the state decided to rely solely on scores on the SAT. This decision resulted in male students receiving disproportionally more scholarships than females, and it paved the way for the plaintiffs to bring the nation's first challenge to an educational testing practice under Title IX of the Education Amendments of 1972. At the time, Title IX stated as follows:

> No person shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any . . . education program or activity operated by a recipient which receives Federal financial assistance. (Office of Civil Rights, 2021)

In addition, the plaintiffs successfully argued that although the legislative intent of the scholarships was to reward high school achievement, the SAT had not been validated to measure high school achievement. The court found in favor of the plaintiffs and issued an injunction prohibiting NYSED from using the SAT as the sole criterion in awarding merit scholarships. Johnson (1979) noted several reasons for limiting reliance on test scores as the sole criterion in selection:

- The relationship between the test score and the construct that is to be measured may not be the same for all groups.
- Many factors, internal and external to the test, may operate to attenuate test scores for some groups of the test-taking population, and the effects of these factors must be considered when using the test score as a criterion for selection, for example, when making decisions about granting a diploma after a course of study or promotion to the next grade versus retention.
- The validity or appropriateness of interpretations and uses of test scores may differ for different groups in the test-taking population.

All of the aforementioned cases were litigated in the name of fairness and equity in testing and assessment and highlighted the importance of safeguarding valid interpretations and uses of assessment results. Most of these cases were litigated during the height of the civil rights movement and shortly thereafter. It is therefore reasonable to assume that similar cases may be brought before the courts again as sensitivities related to diversity, equity, inclusion, fairness, justice, and accessibility in testing and assessment become more prevalent.