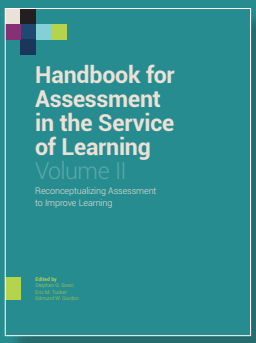
Personalizing Assessment for the Advancement of Equity and Learning

Randy E. Bennett, Eva L. Baker, and Edmund W. Gordon

UMassAmherst
University Libraries

Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker



© 2025 by Randy E. Bennett, Eva L. Baker, and Edmund W. Gordon

The Open Access version of this chapter is licensed under a Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License (CC-BY-NC-ND 4.0).

ISBN: 978-1-945764-33-2

Suggested Citation:
Bennett, R. E., Baker, E. L., & Gordon, E. W. (2025). Personalizing assessment for the advancement of equity and learning. In S. G. Sireci, E. M. Tucker, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume II: Reconceptualizing assessment to improve learning.* University of Massachusetts Amherst Libraries.

Personalizing Assessment for the Advancement of Equity and Learning

Randy E. Bennett, Eva L. Baker, and Edmund W. Gordon

This chapter has been made available under a CC BY-NC-ND license.

Abstract

The purpose of this chapter is to discuss two of the *Principles for Assessment in the Service of Learning* as they affect the design and use of assessments intended to personalize learning for greater impact and equity:

- Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.
- Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

Our discussion begins with background that gives the context for why these principles are needed. That background is followed by a section on some of the various strategies that might be used to achieve the two principles, as well as related design and application issues. The strategies explored are necessarily aspirational in that the state-of-the-art does not yet support their realization at scale. In the third section, we discuss sources of evidence related to quality. We close the chapter with overarching remarks.

Background

In many countries around the world, dramatic population shifts have resulted in unprecedented levels of heterogeneity. Termed "superdiversity" by Vertovec (2023), this condition has come in significant part from immigration, as well as from such other sources as differential birth rates and intermarriage. Concerning race/ethnicity, for example, the U.S. population has evolved from 80% Caucasian in 1980 to 59% in 2021 (Hobbs & Stoops, 2002; US Census Bureau, 2021), a doubling in the proportion of people of color from approximately 20% to 40%. As of 2018, the U.S. public-school population was 53% students of color (National Center for Education Statistics, 2020, Tables 203.70). Throughout Europe, as well as in other parts of the world, countries are experiencing similarly dramatic demographic shifts (Vertovec, 2023).

Many human characteristics are involved in these demographic shifts. Although the most visible might be race/ethnicity, heterogeneity in many locales is also growing along other dimensions. These dimensions include socioeconomic class, immigrant status, native language, gender identity, exceptionality, and sexual orientation. Greater levels of cultural heterogeneity inevitably arise as a consequence of these various kinds of diversity.

Cultural diversity, however, poses a significant challenge (and significant opportunities) for educational assessment in that assessments are cultural artifacts (Greenfield, 1997; Solano-Flores, 2019). Those artifacts presume certain modes of expression, ways of knowing, linguistic structures, and forms of representation. Although populations have changed dramatically, our assessments have remained relatively fixed over time, creating a misalignment. To illustrate, the most obvious change in operational testing programs over the same 1980 to 2021 period in which the US experienced a dramatic shift in racial/ethnic composition was the transition from paper to computer delivery, not a very deep change at all. Classroom assessment practice also has not kept pace in that the demographic composition of public-school teachers is very different from that of students. In general, relatively little investment appears to have been made by federal or state governments in helping a largely Caucasian and middle-class teaching force adjust their assessment and instructional methodologies to the high levels of cultural diversity with which many teachers are confronted.

Noting this growing need, various proposals have been made for accommodating variation in learner populations. Those approaches include culturally responsive assessment (Hood, 1998; Lee, 1998), justice-oriented assessment (Randall, 2021), socioculturally responsive assessment (Bennett, 2023), antiracist assessment (Inoue, 2015), culturally sustaining classroom assessment (Lyons et al., 2021), and universal design for assessment (Ketterlin-Geller, 2005). These approaches may vary along such dimensions as definition, goals, principles espoused, populations targeted, underlying literatures, and intended setting (Bennett, 2025). Even so, most such approaches have at least one shared idea: Designing assessment for the social, cultural, and other relevant characteristics of individuals and the contexts from which they come.

Among the many issues these approaches raise is how to conceptualize diversity, as that conceptualization dictates how to respond in an assessment context. Crenshaw's (1989) insight was that individuals might have multiple identities that could lead to multiple marginalizations. In the context of completing government forms, Rogoff (2003; Rogoff & Angelillo, 2002) described the "box" problem as perceiving groups as homogenous and mutually exclusive when, in fact, they are not. Lee and colleagues (Lee et al., 2020) observed that individuals belong to multiple, not single, cultural groups. Building on these ideas, Vertovec (2023) advocated for "category +," the idea that an individual might primarily affiliate with one group, but secondarily identify with several others. Vertovec argued that such multiple affiliations increase tolerance as one creates more heterogenous connections. Gordon et al. (2019) and Darling-Hammond (personal communication, December 19, 2022) push this idea further still by arguing for an individual-differences or human-variance approach, whereby every learner brings a unique sociocultural identity to learning and assessment.

These conceptualizations of diversity, especially those of Gordon et al. (2019) and Darling-Hammond (personal communication, December 19, 2022), argue strongly for adaptation. Adaptation in educational assessment, and in education more generally, is not new. Adjusting to competency level dates at least to the Stanford-Binet Intelligence Scales, in which the examiner adjusted the difficulty of questions asked based on the answers given by the examinee. That idea appeared in the 1980s in a more theoretically grounded form as computerized adaptive testing (Ward, 1988). Separately, real-time, theory-based adaptation to an individual's domain competencies characterized intelligent tutoring systems (Sleeman & Brown,

1982). Both adaptive tests and adaptive learning systems are now widespread. Other types of adaptation also are currently found in operational assessment programs, including adaptation to exceptionality (in every major test), native language (in many state accountability tests and every international group-score assessment), and interests (various Advanced Placement examinations allowing problem choice or even problem design).

The goals of adaptation in these programs and in the aspirational assessments we envision are described in the section on Evidence of Quality. For now, it is sufficient to note that those goals extend to enhancing such factors as identification with the assessment, engagement and motivation to perform, test performance, learning, perceptions of the test as fair, cultural identity, and meaningful interpretation of results. Some of these factors (e.g., engagement and motivation to perform, cultural identity) will be impacted far more significantly by influences outside of the assessment (e.g., family, community). The goal of assessment adaptation is enhancement, however small such enhancement may prove to be relative to other influences.

Strategies for Adaptation

In this section, we consider approaches to adapting learning and assessment to individual differences, including to the needs and desires of learners. Like many good ideas, it is easier said than done.

Central to the consideration of adaptation are the following questions:

1. What are the purposes of adaptation?
2. What is to be learned and/or assessed?
3. What is to be adapted?
4. On what basis does the adaptation occur?
5. How does adaptation interact with complex learning and assessments?
6. What evidence can be developed in support of adaptation?

Adaptation can be conceived of as multiway matches or interactions among student individual difference variables, instruction, and assessment components or measures that react to interactions. In other words, employed adaptations should improve the performance of target students.

There is considerable history and present practice using measures of student prior knowledge or ability as the basis of adaptation. In addition, there is persistent interest in variables such as learning style and student preferences, although their effects are not well supported (DeBruyckere et al., 2015; Hattie, 2023). Indeed, the lack of positive findings may shift our strategies for applying adaptive approaches from those with evidence of effectiveness to those which "do no harm." We will discuss the implications of lack of convincing, positive data in the light of the clear need for rapid progress in adaptation and policies supporting the evaluation of learners. As we step through the process, both pitfalls and opportunities will emerge.

1. What are the purposes of adaptation?

Our starting point is that adaptation, or the provision of differential options or experiences for learners, is intended to increase for learners both the effectiveness and the equity in the learning-assessment process (Buzick, Casabianca, & Gholson, 2023). Fairness and equity used to be framed as providing the exact same instruction and assessments to each student. One can consider the early designs of standardized test administration, including timed performance, as the pursuit of uniformity, the hallmark of fairness. Everyone was to be given the same experience. However, as our learning about variations in cultural and personal experiences changed (See for example, Wenger, 1998), so have our formulations of fairness. Instead of thinking about equity as providing uniform exposure in highly specified conditions, we have more recently focused on how elements of the situation and how the content of the material may be interpreted differently and what can be done to assist different respondents to display their most positive or representative level of expertise. Adaptation in learning situations may involve wide-ranging differences in sequence, repetition, complexity, and feedback. In assessment that is a part of learning, adaptation of timing, resources, and content are more easily included in instructional variations based on sequences of learners' answers, errors, and timing. In a formal assessment setting, which typically is far shorter than the instructional experience, adaptations present greater design challenges.

2. What is to be learned and/or assessed?

Assessments vary in many ways, including their purposes, tasks, domains, and variety. It should be expected that specific features of assessments (e.g., multiple steps, reliance on domain knowledge) will trigger varied student reactions resulting in performance differences. For example, outcome measures differ between and within their designs in terms of expected cognitive demands on the assessment. Methods to classify types of cognition in learning or on examinations are both numerous and varied. A major classification branch on assessment arises from the venerable works of Bloom and his associates (1956) as updated by Anderson and Krathwohl (2001). These approaches examine the assessment and its items to characterize their cognitive demands. Other classification systems focus more directly on the learning process, such as those generated by Gagné and Medsker (1996), Merrill (2009), and Webb (1997). In simplified terms, all these systems subdivide learning processes and outcomes into at least two categories: (a) access and understanding knowledge, and (b) applications of knowledge and other thinking processes (See also Hattie et al., this volume).

A third area of great importance, but limited research, is how knowledge and thinking processes may be applied differently to situations ranging from the familiar (those used in the instructional examples) compared to tasks embedded in previously unencountered contexts or problems. In educational psychology, the movement from familiar to unfamiliar situations in performance is called "transfer" (See Perkins & Salomon, 1992). Clearly, when test situations are unfamiliar to students, such as learners from diverse backgrounds, we are functionally adding the requirement of transfer to their learning objectives, as compared with students who may, because of their background, be already accustomed to the contexts and conditions of the assessment.

Research on transfer does in fact indicate that the presentation of unfamiliar contexts and conditions generally diminishes performance, giving one additional explanation for persistent lower performance of some groups. Even seemingly unimportant differences in assessments, such as the use of synonyms for vocabulary, can have an effect (Abedi & Ewers, 2013). The phenomenon of transfer, then, leads to the premise that students who have less broad, or different experiences, will likely perform less well when they are less accustomed to variations in assessment and learning situations.

How to change this situation is not clear. The willingness to engage in unfamiliar tasks may be encouraged by caregivers or the community (See Gordon, 1984), but by itself such encouragement is unlikely to reduce the difficulty of such tasks for the learner. As Gordon (personal communication, August 1, 2024) has asserted, to be successful minoritized students must learn the intellective habits of mind and curricular content dictated by the dominant group, however unfair that might seem. Advocates of culturally attuned approaches to teaching (Gay, 2002, 2018; Ladson-Billings, 1995; 2021; Moll, Amanti, Neff, & Gonzalez, 1992) argue for using the knowledge and identity brought from home and community as conduits to learning such dominantly valued content. Advocates also argue for broadening that dominantly valued content. Broadening has in fact occurred in some states (Bellamy-Walker, 2022; Mays, 2021; NYSED, n.d.), though in other states the opposite appears to be happening (Cavanaugh, 2023).

Thus, despite the logic of making an assessment more familiar to the learner (again, to avoid unintended transfer load for diverse students), the issue is a complex one. As the underlying logic and transfer research might suggest, the resulting scores may not generalize to performance in other contexts, particularly the contexts valued by the dominant group. Moreover, the evidence base in support of adapting instruction and assessment is neither large nor overwhelmingly positive. Given that many research studies used standardized, general assessments of constructs rather than ones more attuned to the types of transfer that might be expected from what has been specifically emphasized in instruction, a new wave of studies of adaptation is obviously needed and is partially described in a subsequent section.

3. What is to be adapted?

Let us begin with fundamental questions about the premises and methods of developing approaches that adapt assessments to learner needs. We then proceed to address approaches that may offer substantial inroads to the solution of a heretofore intractable problem. Our initial considerations focus on what interventions may be applied, while the second set of concerns attends to how we might go about their implementation.

Looking for interventions that help learners with different measured abilities, needs, or preferences has occupied educational psychologists for decades. For instance, in their landmark study investigating ability-treatment interactions in learning, Cronbach and Snow (1977) concluded that the lack of positive findings

may be partly attributed to the character of dependent or outcome measures used in the study (pp. 170–171). Cronbach and Snow suggested that where studies use well-defined outcomes, performance is likely to be more positively affected by adaptations than when more general measures are used. They cited Atkinson's (1972) micro-adaptation as an example. Their research is among many that have sought to discover reliable interactions among learner background variables, attributes of instruction, and assessment systems that positively affect learner performance.

In both learning and assessment, we assume that sets of learners need different experiences, tasks, cues, and sequences to assist their achieving desired performance. Learning sequences and extended examinations involve content or domain knowledge, situations, and tasks, the last of these including requirements, response formats, and stringency levels of judgment. Various terms have been used to describe adjusting testing experiences to support the respondent. Perhaps the most used has been the term "accommodations," generally supporting performance by relaxing time limits, providing additional cues such as glossaries, or enabling opportunity to review items of content (Abedi & Ewers, 2013).

The target of adaptation can be addressed in two large and somewhat overlapping sets of variables. The first focuses on components in the learners' repertoire that clearly relate to the desired performance. These components may reflect essential subcomponents needed for successful performance. The second type of adaptation target is indirectly related, where the variable serves as a moderator to more directly related behaviors.

Variables for Adaptation

The prior knowledge of learners serves as a frequent way to determine adaptation. This approach is more situation and domain specific than measures of general ability considered by Cronbach and Snow (1977). In Hattie's syntheses of meta-analyses of research (2023), he reports a weighted effect size of .96 for prior ability and intelligence, and an effect size of .73 for prior achievement of learners. These findings suggest the potential efficacy of identifying learner differences that are connected directly to identifiable components of instruction and assessment. In fact, such prior knowledge dominates the field of adaptation, whether presented in computer platforms, as lessons or learning games, or in tests. Employed as a strategy in early instances of programmed instruction (Pressey, 1950), prior

knowledge displayed in initial performance has been at the core of computer adaptive testing (CAT) for decades (Weiss, 2011). CAT operates by zeroing in on where the learner is in a framework of assessment items. It uses initial knowledge to bracket and then confirm performance levels. While widespread, a major function of CAT is to make the testing process more accurate and efficient, rather than to support learner performance.

Other direct measures, perhaps observed during the learning/assessment process in computer learning systems, include speed of response and pattern of errors on the tasks of interest. These measures may be created top-down, as a modification of rule-based design, or inferred from deep learning or other bottom-up processes. Using the fundamental design of intelligent systems, described by Mislevy (2018) and by Shute and others (2015), individual response data reside in a student model which is updated as continued responses are made. To determine the best next action, the system may compare the learner's patterns to either expert models or to the repository of patterns obtained from other students. Based on these analyses, a recommendation for the next item or sequence is generated and implemented. This approach can result in varied numbers of responses elicited from the student and shortening or lengthening the learning/assessment experience as needed.

The difficult fact is that students who have been found to lack elements essential for further learning, for instance, mastery of a particular math procedure, will need to acquire it in order to undertake the desired task. Catching up without taking more time is a difficult proposition. One can reflect on approaches that may allow students to overcome missed skills or knowledge without spending considerable portions of available time, where such time is relatively fixed. Certainly, familial motivation is an important factor, but not one that is easily amenable to change (Gordon, personal communication, 2024). One strategy that received strong moral support was advocated by Oakes (1992), who wished to "detrack" learning by placing students in heterogeneous groups in classrooms. Overall effects for detracking are small (ES = .09) and comparable for findings in tracked classroom (ES = .09) (Hattie, 2023). In tracked environments, lowest level students did the worst, perhaps because of the catching-up process. Conversely, in detracked studies, it was lower performing students who were most positively affected.

Most intriguing, perhaps, are findings about approaches intended to increase the student's agency in learning, such as self-judgment and reflection (ES = .69), metacognitive strategies (ES = .52), and self-regulation (ES = .52; Hattie, 2023, p. 349). Brummelman and Sedikides (2023) also argued for approaches to animate students' agency. Nonetheless, student control over instruction does not rank high on the meta-analytic summaries of effects of learning (ES = .04), but ranks somewhat higher in meta-analyses of motivation (ES=.30; Hattie, 2009, p. 193). We can posit a few reasons. First, it may be possible that the options offered are not functionally different. The option to write about "how I spend my free time" versus "my favorite summer vacation" could appear to be identical to a given learner. Developing choices that connect specifically to varied aspects of different learner's experiences is a non-trivial task. A second concern is that even with reasonably different choices, the learner may not know how to select the best personal choice. In this case, the teacher or instructional system should be enabled to give useful (validated) instruction on how to make choices. This instruction might include encouraging the learner to answer a series of questions about the experience with the choice, the need for further information or support, and their ability to recognize a good choice. Such instruction may well be conducted collaboratively.

Beyond the ability to choose procedures or options most likely to result in improved learning, at some point, choice should also include goals and objectives that supplement or depart from the uniform curriculum.

Indirect Variables for Adaptation

Indirect approaches to adaptation focus on variables or experiences that surround and may interact with the learning-assessment process. Among the many examples here are efforts to reduce test anxiety, a variable that has a reliable negative relationship to performance (Rana & Mahmoud, 2010). Hattie describes a .30 effect size on achievement in meta-analyses of test anxiety reduction (Hattie, 2009, pp. 49–50) by providing respondents with choices to control the assessment environment. Other indirect approaches operate on presumably improving motivation through use of engaging or amusing episodes, pictures, or content that is likely to interest the learner. Mayer (2014) and others before him have cautioned that such options, unless carefully designed, can distract the learner, and reduce their likelihood of demonstrating their best performance. Moreover, options that take more time to complete can reduce performance when time limits are imposed.

Another major area of interest is motivation as an intervening variable affecting performance. Bandura (1977) noted positive emotional and physiological states support learning. In 1982, he added comfort and experience in the cultural setting, which can assist learners to perform, a point amplified by the work of Duran (2022). Motivation may be considered as arousal or seeking to engage in learning. Technological approaches to measuring motivation and arousal as processes, using measures of eye movement, pupil dilation, and other sensor-based approaches, have been well summarized by Plass and Kalyuga (2019).

There are also conceptual and practical questions related to these approaches as strategies for equity. A principal concern is that many of these variables often show main effects on learning and performance; that is, they help everyone to some degree. If equity is framed as equal opportunity, this concern should not be central. But in discussions underlying approaches to diverse students, an assumption is often made about differential and potentially accelerated progress, in which case main effects are not as desirable as interactions (Sireci, Scarpati, & Li, 2005).

4. On what basis does the adaptation occur?

Deciding on who gets adaptation may depend on performance on a pre-instructional or initial assessment indicator (or indicators) obtained prior to encountering the bulk of the learning or assessment. In many assessment systems, algorithms are used to capture and modify the learner's progress in the episode of assessment and may well include earlier patterns of performance. Different learners may receive adaptations on different schedules. It is less clear whether there are empirically verifiable cut scores for the decision(s) to adapt performance.

This discussion raises a significant conundrum. How do we avoid stereotyping students based on assumptions about their cultural experiences, language status, and socioeconomic characteristics? Without an accurate picture of the learner, how do we maintain our interest in, and adaptation to, the individual differences the student has, specifically those that may be independent of, or diverge from, those differences generally ascribed to the population group? Consider if the construction of a model of genomic analysis of learning, somewhat like the LEARNOME analogy by Baker (Baker et al., 2002) or that of MIT (2024), might help. Here detailed attributes of, and progress in, the learner's performance and other measured skills are integrated in a description of capability. The complexity of personalization in such models may result in interventions that assist students beyond the use

of group stereotypes. We can make progress on effective adaptation using our present approaches, but serious attainment of equity is not likely without identifying the multitude of critical knowledge, skills, and attributes of students, as well as their interactions.

5. What about complex assessments?

To address the question of adaptation for learning more complex behaviors, and for assessment that embeds learning (e.g., dynamic assessments), we start with the literature in learning. Most stand-alone assessment situations may be of relatively short duration. Students may also recognize the stakes pertaining to their performance. In complex assessments, with an extended process or product, like an oral science report on a set of experiments, an essay based on research, or a process, such as a musical performance, the student manages several individual and interacting elements to complete the task. In the conduct of such complex assessments, support may be given to assure that required prerequisites (such as knowing the right equation, identifying a tractable problem, or accessing suitable examples) are provided. That support may either take the form of additional resources or assets provided during the assessment, or feedback given on adequacy of the learner's formulation of a needed step. Clearly, there is a blurring of the boundaries between assessment and learning, the integration of which for many observers is not a problem. On the other hand, if the assessment is formulated so students are intended to display competence on a generally comparable set of requirements, and if the claimed comparability is the basis of high stakes decisions, then any adaptation that substantially compromises comparability may be out of the question. Rather, we propose thinking about the adjustments here as tuning the assessment to the learner, at once a more precise and less burdensome approach. One option that may meet these desires is giving choices to the learner of content, emphases, and presentation, despite the ambiguous data in its support (Powers & Bennett, 1999).

We have reviewed steps important to the determination of approaches to adaptation including why, what, and how. In the next section we consider the questions that might be used to frame our thinking theoretically about the evidence of adaptation quality (question 6, discussed earlier).

Evidence of Quality

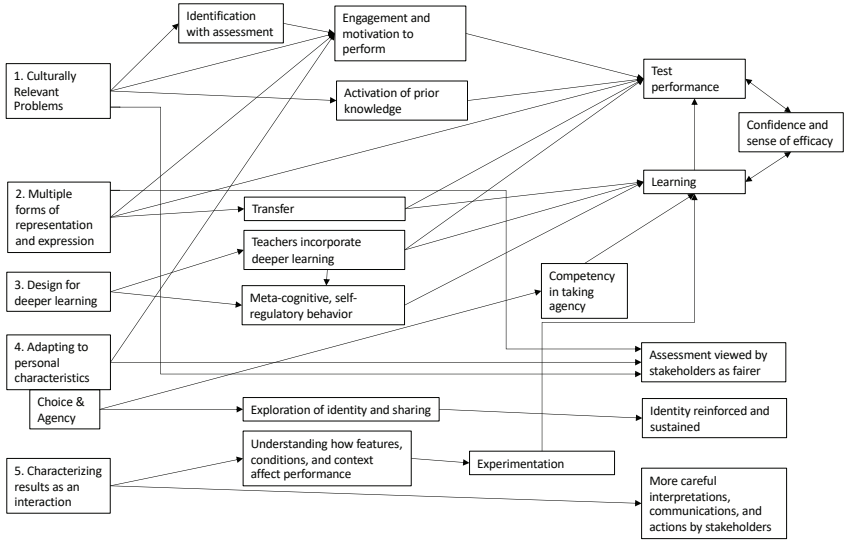
To date, the evidence on adaptation to personalize learning and assessments is scattered, based on partial information, small sample trials, and particularized settings. How might one determine whether an assessment that was personalized for the advancement of equity and learning was effective? Bennett (2023) proposed a theory for socioculturally responsive assessment that offers one approach to addressing this question. The theory is built from five assessment design principles drawn from multiple literatures, including those on the teaching of traditionally underserved students, the assessment of such students, and the learning sciences. The principles are as follows:

1. Present problem situations that connect to, and value, examinee experience, culture, and identity because individuals are more likely to show what they know and can do in familiar versus unfamiliar situations;
2. Allow for multiple forms of representation and expression in problem stimuli and in responses, providing more equitable access to question content and to paths for response;
3. Promote instruction for deeper learning through assessment design to increase the likelihood of exposing students in lower-performing schools to such learning;
4. Adapt the assessment to student characteristics, including by offering choice of problems and agency in contributing to problem definition; and
5. Represent assessment results as an interaction among what the examinee brings to the assessment, the types of tasks engaged, and the conditions and context of that engagement, thereby giving a more nuanced depiction of performance.

As should be clear, except for promoting deeper learning through assessment design, each of these principles directly implies personalization: presenting problems that connect to examinee experience, culture, and identity; allowing forms of representation and expression suited to the examinee; adapting the assessment to student characteristics; and framing results to reflect the unique intersection of factors brought by the examine and by the assessment. Designing the assessment to promote deeper learning is aimed at encouraging adaptations that maintain rigor.

Figure 1 shows how these design principles work together theoretically to produce their intended effects. In the Figure, the theory is represented as a network of empirically testable propositions emanating from the design principles. On the left are the five principles. To their right are intermediate and ultimate outcomes (i.e., the postulated effects of designing and implementing an assessment according to those principles). In terms of the theory, a personalized assessment is effective to the degree that a preponderance of the posited propositions are supported by the empirical evidence. Next, the principles are listed, in shortened form, each followed by a discussion of the propositions, along with the types of studies that might be used to evaluate them. Because it is not particular to personalized assessment, principle #3 (promote instruction for deeper learning …) is not discussed.

Figure 1.
An initial theory of socioculturally responsive assessment



Note. Adapted from "Toward a theory of socioculturally responsive assessment" (p. 97), by R. E. Bennett, 2023, *Educational Assessment, 28*(2). Copyright [2023] by Educational Testing Service. Adapted by permission.

Principle 1: Present problem situations that connect to examinee experience, culture, and identity

The propositions emanating from this principle posit that a learner's identification with an assessment will increase when the assessment incorporates problems connected to their cultural identity, background, and lived experiences. Increased identification should in turn promote higher levels of engagement and motivation to perform. In addition, the incorporation of culturally relevant problems should help to activate prior knowledge, causing students to perform better than on problems that do not make such personal linkages. Higher levels of performance should raise confidence and bolster sense of efficacy, thereby enhancing learning and future test performance. Lastly, stakeholders should be led to feel that the assessment is fairer.

This proposition set might be tested through studies that experimentally manipulated the cultural relevance of problems for individuals, examining the degree to which aligned problems resulted in the described effects: for examinees, increased identification, engagement and motivation to perform, activation of prior knowledge, and performance; for stakeholders, improved perceptions of fairness.

Principle 2: Allow for multiple forms of representation and expression in problem stimuli and in responses

Students should be better able to show what they know and can do when they are offered multiple forms of representation (e.g., essay, bulleted list, drag-and-drop graphical form) and allowed alternate modes of expression (e.g., spoken, written, sign). Such affordances should cause increased student performance and perceptions among stakeholders that testing is fair. When assessment incorporates problems that encourage making deep-structure connections among representational forms and expressive modes, subsequent transfer of learning should be enhanced and test performance improved.

The propositions associated with this principle might be tested through studies that experimentally manipulated the allowance of single versus multiple forms of representation and single versus multiple forms of expression. For example, in the case of a science content assessment, students might be randomly assigned to a condition that permitted choice of representational form (e.g., brief essay, bulleted list, drag-and-drop graphical form) versus one that required responding only via a brief essay. Along with the main effects on engagement, motivation,

and performance, in a sample large enough to disaggregate demographically, the group-by-condition interaction could also be evaluated. Such studies should evaluate whether student-made choices actually benefited performance or are made on more superficial considerations, for instance, to shorten testing time (See Powers & Bennett, 1999, for a review of research on choice).

Principle 4: Adapt the assessment to student characteristics

The propositions linked to this principle concern the postulated effects of adjusting to personal characteristics. Such adaptation should better align the assessment with student interests, cultural identity, background, and prior knowledge than does a traditional test, thereby causing stakeholders to feel the assessment is fairer. Higher levels of motivation and engagement with the test should also result, thereby enhancing performance. With teacher guidance in making similar choices over the school year, permitting choice should aid the development of student competency in taking agency. That competency should, in turn, impact learning positively. Moreover, when allowing agency motivates examinees to examine their identity and share those explorations, identity should be reinforced and sustained. Especially for minoritized students, the positive effects of adaptation should increase with the degree to which the assessment meaningfully takes personal characteristics into account.

Among other studies, members of this proposition set might be evaluated by asking stake holders to rate the fairness of hypothetical examples of adaptation to particular student characteristics, as well as non-adaptive examples. In addition, asking students to rate engagement and motivation to perform, after taking personalized versus non-personalized tasks designed to measure the same content, would be informative, as would evaluating any performance differential and its direction. Questions of comparability might also be evaluated by correlating scores from both traditional and personalized assessments with measures of criterion performance, and by structural equation modeling approaches.

Principle 5: Represent assessment results as an interaction among the examinee, the tasks, and the conditions and context

The propositions implied by this principle postulate that examinees, teachers, the public, and policy makers will more carefully interpret, communicate about, and act on assessment results if those results are represented to them as an interaction among examinee characteristics, the features of tasks, and the conditions and context of the assessment. Because the notion of fairness has been embedded in uniformity, strong evidence of the utility of adaptations should be required. In this regard, more careful interpretation implies viewing results as closely tied to the task types, conditions, and contexts that characterized the assessment. When results are thought of in this way, students should be more likely to make connections between how task features, conditions, and contexts affect their performance. Additionally, teachers and students should be led to explore adjustments to these factors to see how they might impact learning and improve test performance. Studies that provide a treatment to teachers and students designed to sensitize them to this interactional perspective, followed by cognitive interviews and classroom observations, might help uncover whether the postulated effects emerge.

Conclusion

Our thesis has been simple, that assessments (and learning) should be designed and implemented to provide flexibility and adaptation to individual differences. This orientation makes demands on many groups. Some of the burden falls on test developers who must deal with the complexities of implementation. Teachers' engagement will also be essential if students are to be given differential types of learning and assessments, in particular to equally value those types and communicate that perspective to students. Students are also encouraged through metacognitive and self-regulated thinking to become more thoughtful about the choices they make in assessment contexts (Bembenutty, this volume). Policy makers and parents need preparation and examples so they can accept assessments that differ among students.

A tremendous burden falls upon researchers and the schools and districts which must help in generating the evidence. Evidence needs to be generated to explore the utility of the model systematically and to provide guidance for its revision, as needed. Teacher and student participants need to be available, across individual

and group variations, subject matter domains, skills, and age ranges. Perhaps the evaluation strategy should start with an age range, such as that found in middle-school, and a particular subject matter. If positive evidence is found, a second step might be to evaluate the effects in other domains and other age levels. Whatever the particular research strategy, it is clear that research into the major premises of this work will ultimately need to be supported by government, commercial test developers, purveyors of instructional systems, and teacher development organizations. The problem of endemic differential performance must be finally addressed, and it will take more than a village to do so.

References

Abedi, J., & Ewers, N. (2013). *Accommodations for English language 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

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete ed.). Longman.

Atkinson, R. C. (1972). Ingredients for a theory of instruction. *American Psychologist, 27*(10), 921–931.

Baker, E. L., Sawaki, Y., & Stoker, G. (2002, April 1–5). *The LEARNOME: A descriptive model to map and predict functional relationships among tests and educational system components* [Paper presentation]. Annual meeting of the American Educational Research Association, New Orleans, LA, United States.

Bandura, A. (1977). *Social learning theory.* Prentice-Hall.

Bellamy-Walker, T. (2022, January 18). *NJ mandates teaching Asian American and Pacific Islander history in schools*. NBC News. https://www.nbcnews.com/news/asian-america/nj-mandates-teaching-asian-american-pacific-islander-history-schools-rcna12637

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Bennett, R. E. (2025). A descriptive review of culturally responsive, socioculturally responsive, and related assessment conceptions. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 11–28). Routledge. https://doi.org/10.4324/9781003435101

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* David McKay.

Brummelman, E., & Sedikides, C. (2023). Unequal selves in the classroom: Nature, origins, and consequences of socioeconomic disparities in children's self-views. *Developmental Psychology, 59*(11), 1962–1987.

Buzick, H. M., Casabianca, J., & Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educational Measurement: Issues and Practice.* https://doi.org/10.1111/emip.12551

Cavanaugh, S. (2023, February 27). State restrictions like Florida's on curriculum are having an effect nationwide. *EdWeek Market Brief.* https://marketbrief.edweek.org/regulation-policy/state-restrictions-like-floridas-on-curriculum-are-having-an-effect-nationwide/2023/02

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum, 1*(8) 139–167. http://chicagounbound. uchicago.edu/uclf/vol1989/iss1/8

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

De Bruyckere, P., Kirschner, P. A., & Hulshof, C. D. (2015). *Urban myths about learning and education.* Academic Press.

Durán, R. (2022, November 1–2). *Cultural context and affirmative pedagogies* (Facilitator). Extending His Reach: Promoting the Educational Legacy of Edmund W. Gordon, hosted by the CRESST, UCLA School of Education & Information Studies, and Gevirtz School of Education, UC Santa Barbara, Los Angeles, CA, United States.

Gagné, R. M., & Medsker, K. L. (1996). *The conditions of learning: Training applications*. Harcourt Brace.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106–116.

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice* (3rd ed.). Teachers College Press.

Gordon, E. W., Boykin, A. W., Armour-Thomas, E., & McCallister, C. (2019). *Human variance and assessment for learning*. Third World Press Foundation.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*(10), 1115–1124.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Hattie, J. (2023). *Visible learning: The sequel. A synthesis of over 2,100 meta-analyses relating to achievement.* Routledge.

Hobbs, F., & Stoops, N. (2002). *Demographic trends in the 20th century: Census 2000 special reports* (CENSR-4). US Census Bureau. https://eric.ed.gov/?id=ED477270

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

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. The WAC Clearinghouse; Parlor Press. https://doi.org/10.37514/PER-B.2015.0698

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2).

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465–491.

Ladson-Billings, G. (2021). *Culturally relevant pedagogy: Asking a different question.* Teachers College Press.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education, 67*(3), 268–279.

Lee, C. D., Meltzoff, A. N., & Kuhl, P. K. (2020). The braid of human learning and development: Neurophysiological processes and participation in cultural practices. In N. S. Nasir, C. D. Lee, R. Pea, & M. McKinney de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 24–43). Routledge.

Lyons, S., Johnson, M., & Hinds, B. F. (2021). *A call to action: Confronting inequity in assessment.* Lyons Assessment Consulting. https://www.lyonsassessmentconsulting.com/resources.html

Mayer, R. E. (2014). Incorporating motivation into multimedia learning, *Learning and Instruction, 29*, 171–173.

Mays, M. (2021, October 14). California students will have to take ethnic studies to get a diploma. *Politico*. https://www.politico.com/news/2021/10/14/california-students-ethnic-studies-diploma-515972

Merrill, M. D. (2009). First principles of instruction. In C. M. Reigeluth & A. Carr (Eds.), *Instructional design theories and models: Building a common knowledge base* (Vol. III). Routledge.

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

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into Practice, 13*(2), 132–141.

National Center for Education Statistics (NCES). (2020). *Digest of education statistics: 2020 tables and figures.* US Department of Education. https://nces.ed.gov/programs/digest/d20/tables/dt20_203.70.asp

Oakes, J. (2005). *Keeping track: How schools structure inequality.* Yale University Press.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 425–441). Pergamon.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review, 31,* 339–359.

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. doi:10.1207/S15324818AME1203_3

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology, 29*, 417–447.

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

Rana, R., & Mahmood, N. (2010). The relationship between test anxiety and academic achievement. *Bulletin of Education and Research, 32*(2), 63–74.

Rogoff, B. (2003). *The cultural nature of human development.* Oxford University Press.

Rogoff, B., & Angelillo, C. (2002). Investigating the coordinated functioning of multifaceted cultural practices in human development. *Human Development, 45*, 211–225.

Shute, V. J., D'Mello, S. K., Baker, R., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224–235.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457–490.

Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. Academic Press.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The Matrix of evidence for validity argumentation. *Frontiers in Education, 4*(43). https://doi.org/10.3389/feduc.2019.00043

US Census Bureau. (2021). *Quick facts: United States*. https://www.census.gov/quickfacts/fact/table/US/PST045219

Vertovec, S. (2023). *Superdiversity: Migration and social complexity.* Routledge.

Ward, W. C. (1988). The College Board Computerized Placement Tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 271–282.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education* (Research Monograph Number 6). CCSSO.

Weiss, D. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1–27.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge University Press.