*ets research institute

JANUARY 2026
RM–26-01

RESEARCH MEMORANDUM

A Preliminary Research and Evaluation Agenda for Personalized Assessment in the Service of Equity

AUTHORS

Randy E. Bennett, Jesse R. Sparks, Burcu Arslan, Blair Lehman, Sandip Sinharay, and Diego Zapata-Rivera

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

A Preliminary Research and Evaluation Agenda for Personalized Assessment in the Service of Equity

Randy E. Bennett, Jesse R. Sparks, Burcu Arslan, Blair Lehman, Sandip Sinharay, & Deigo Zapata Rivera

ETS Research Institute, Princeton, New Jersey, United States

January 2026

Suggested citation: Bennett, R. E., Sparks, J. R., Arslan, B., Lehman, B., Sinharay, S., & Zapata-Rivera, D. (2025). *A preliminary research and evaluation agenda for personalized assessment in the service of equity* (Research Memorandum No. RM-26-01). ETS. https://doi.org/10.64634/wjv5e895

Find other ETS-published reports by searching the
ETS ReSEARCHER database.

To obtain a copy of an ETS research report, please visit
https://www.ets.org/contact/additional/research.html

Action Editor: Daniel F. McCaffrey

Reviewer: Patrick Kyllonen and Andrew McEachin

Table of Contents

Abstract

This research memorandum describes a preliminary research and evaluation agenda for personalized assessments, where such assessments are intended to be attuned to the social, cultural, and other relevant characteristics of individuals and the contexts from which they come. The agenda targets the full range of assessment uses—school accountability, national and international assessment, admissions, certification and licensure, and instructional planning. The purposes of the agenda are to guide the theoretical and empirical research and development needed to create personalized assessments and to suggest a means for judging the effectiveness of those instruments.

Keywords: personalized, assessment, agenda, equity, fairness

Introduction

This research memorandum describes a dual-purpose investigatory agenda for personalized assessments that are sensitive to the social, cultural, and other relevant characteristics of individuals and the contexts from which they come. The goal of such assessments is to measure the competencies of individuals, especially from minoritized groups,[1] more validly across the full range of assessment uses—school accountability, national and international assessment, admissions, certification and licensure, and instructional planning. Thus, in contrast to standardized assessments, which try to minimize construct-irrelevant difficulty *on average*, personalized assessment tries to minimize such difficulty for the *individual* (Mislevy et al., 2013).

This agenda has two main purposes. The first purpose is to guide the theoretical and empirical research and development needed to create assessments that differentially adjust to the needs of individuals. The second purpose is to suggest a means for judging the efficacy of those assessments.

In the context of this agenda, a personalized assessment might in theory be designed to operate in one of at least three general ways, each of which may have many variations (Bennett, 2023, 2024). One way is machine-driven. In this approach, test designers select examinee characteristics for artificial intelligence (AI) models to use in adjusting the assessment

content, format, response modality, and other conditions in real time to the individual. A second general approach is examinee-driven, which means the assessment is engineered such that the test taker can decide in varying degrees whether and how to bring their characteristics to the assessment to best depict what they know and can do. The range of admissible characteristics may be restricted by the designer or sponsor (e.g., by offering choice among problems created to appeal to a limited set of personal characteristics or by allowing examinees to design problems within constraints). Such *Advanced Placement*© examinations as the *AP*© United States History test (College Board, 2023c), AP Computer Science Principles (College Board, 2023d), AP Research (College Board, 2023a), and AP Art and Design (College Board, 2023b) exemplify this approach to varying degrees. A final possibility is a combination of these two general approaches. Arslan (2024) described an example in which the examinee indicates their area of topical interest and mathematical problems are then customized in real time to that topical context using generative AI. Common to all the approaches is an assessment designed to the maximum degree possible for the individual rather than for any given demographic group.[2]

A final introductory note is that this agenda focuses on the scientific aspects of personalized assessment, in particular building theory, evaluating design principles, and amassing other knowledge and capability to undergird this form of assessment. The agenda does not take on such related challenges as operational implementation or the politics of equity in education and assessment, although such issues will ultimately need to be addressed if the results of this research are to be applied in practice.

Research and Evaluation Studies

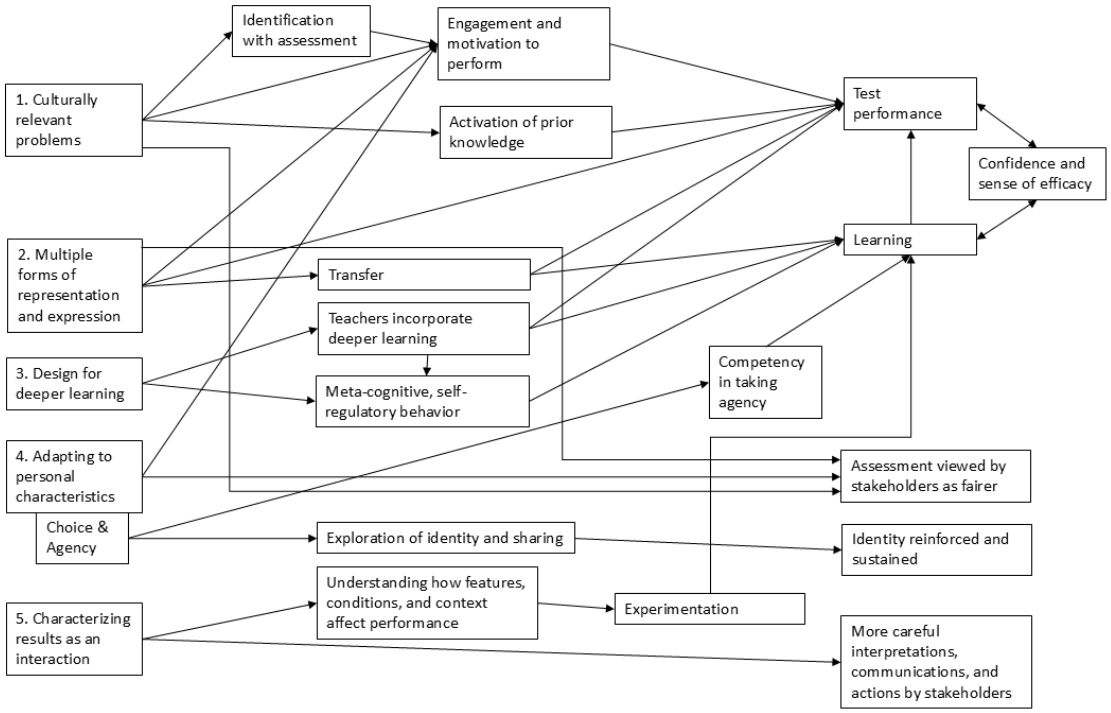
We use as an initial framing the theory of socioculturally responsive assessment proposed by Bennett (2023). That theory represents one way in which personalized assessment in the service of equity might be conceptualized. The theory is intended to partially explain the fact of lower performance of minoritized groups relative to the majority group on standardized tests in terms of causal factors related to test design. (For a detailed explication of other causes, especially opportunity to learn, see Bennett, 2025.) The theory is composed of a network of empirically testable propositions linked to assessment design principles. The theory

and network were derived from a review of multiple literatures, including those on the teaching and assessment of minoritized students and from the learning sciences. The propositions offer a ready starting point for research that will be elaborated over time as propositions are amended and added, thereby refining the theory. Thus, the theory is used to provide a coherent framework for selecting research questions, which in turn motivates a step-by-step procession of illustrative studies to evaluate and refine the theory.

The testing of theoretical propositions is necessarily inseparable from the evaluation of the theory's assessment design principles. This intermingling occurs because the theory's propositions are derived from those principles. Consequently, creating and administering an assessment built from one or more design principles is a test of both the theory and the effectiveness of specific principles as guides to instrument design.

Figure 1 summarizes the theory's network of empirically testable propositions, which constitutes the organizing scheme for the agenda. In Figure 1, the five design principles are given on the left, with each principle's propositions taking the form of an arrow leading to an intermediate or ultimate outcome to the right. The propositions associated with each design principle are described in turn.[3] Following each description are one or more research questions implied by the propositions. After each question, a summary of a study that might address that question is given. The summaries are intended as starting points for interested researchers, who would need to familiarize themselves with the related literature, revise the research question as appropriate, develop a detailed study justification grounded in that literature, and flesh out or otherwise amend the suggested design. With respect to design, many of the summaries propose experiments because they offer the most direct test of the theory's propositions and of a principle's value for assessment design. Where experimental (or even quasi-experimental) design did not seem substantively sensible or logistically feasible, qualitative methods were suggested as a means of offering insight into the issue in question.

Figure 1. An Initial Theory of Socioculturally Responsive Assessment



Note. Adapted from "Toward a Theory of Socioculturally Responsive Assessment" by R. E. Bennett, 2023, *Educational Assessment*, *28*(2), p. 97. Copyright 2023 by ETS. Used with permission.

Principle 1: Present Problem Situations That Connect to, and Value, Examinee Experience, Culture, and Identity

Problems that resonate with the cultural identity, background, and lived experiences of all learners—but especially minoritized ones—are posited to cause increased learner identification with the assessment, thereby promoting engagement and motivation to perform.[4] Such problems should help to activate prior knowledge that builds on the assets these learners bring to school (Gay, 2018; Gonzalez et al., 2005; Ladson-Billings, 2021; Walkington, 2013; Walkington & Bernacki, 2020), causing students to perform better than they would on problems that do not make such connections (Bernacki & Walkington, 2018; Ebe, 2025; Lin et al., 2024; Major et al., 2021; Malda et al., 2010; National Research Council [NRC], 2007, pp. 19, 119, 142; Wang et al., 2025; Zheng et al., 2022). Better performance should

contribute to confidence and a sense of efficacy, which, in a virtuous circle, facilitate learning and test performance, returning to confidence and efficacy. Finally, these problems should lead to perceptions among stakeholders that assessment is fairer.

Study 1

Research Question. Do examinees perceive personalized assessments that maximize the relevance of tasks to be more engaging and motivating than current methods? Do examinees show evidence of greater engagement?

Summary. By virtue of including content that resonates with examinees' background, interests, cultural identity, and lived experience, personalization is posited to be more engaging than traditional (i.e., standardized) assessment methods and, thus, more motivating. This claim could be evaluated by presenting personalization exemplars and asking examinees from a variety of backgrounds to rate the exemplars on the degree to which they might find the exemplars engaging relative to more traditional assessments and whether they would be more or less motivated to perform if they were to take them. Exemplars might be found among the subset of ETS Testlets created to be culturally responsive (O'Dwyer et al., 2023). Ebe's (2010, 2025) cultural relevance rubric might be one pertinent rating scale; an additional useful source might be Evans (2023, Table 1). Using the ETS Testlets, examinees from different demographic groups could rate engagement and motivation to perform with respect to segments of one of the culturally responsive forms versus one of the forms measuring a similar construct but not specifically designed to be culturally responsive.

Another approach could be to administer both culturally responsive and baseline versions of assessment tasks to students (whether in a between-subjects or a within-subjects design) and ask them to rate the assessment they took in terms of engagement and motivation. If students complete the tasks using a computer-based platform that enables logging of interactions with test items, it should be possible to derive estimates of engagement with the tasks to facilitate comparisons across conditions, for example, by using measures of response-time effort (Wise & Kong, 2005; Wang et al., 2025). Such work would complement student ratings and provide an additional source of evidence to address claims about the extent to which cultural responsiveness facilitates engagement in assessment contexts.

Study 2

Research Question. Do personalized assessments that maximize the relevance of tasks activate examinee prior knowledge to a greater degree than traditional assessment approaches?

Summary. Research has suggested that prior knowledge significantly influences both learning and understanding (NRC, 2000). For example, students from different cultural groups were more likely to effectively make sense of science test items when they related the item content and contextual information to meaningful aspects of their lives (Sexton & Solano-Flores, 2002; Solano-Flores & Li, 2009; Solano-Flores & Nelson-Barber, 2001). This study could randomly assign students from one or more demographic groups to personalized assessment versus traditional assessment conditions, then measure the degree to which prior knowledge was activated in each group for each condition (e.g., problem context personalization vs. no problem context personalization). Activation might be inferred through various methods. For example, familiarity with topical vocabulary, in which examinees rate their knowledge level for listed words (Wang et al., 2025), could be compared across personalized versus traditional assessment conditions. Response-time evaluation might be another possibility, in which the tested hypothesis would be that time would be shorter for personalized versus traditional, standardized conditions. These methods could be supplemented by running cognitive labs with a few students from each condition to probe for evidence of prior-knowledge activation (e.g., via think-aloud methodologies).

Study 3

Research Question. Do personalized assessments that maximize the relevance of tasks lead to an increase in examinee performance compared with methods that don't attempt such maximization?

Summary. If personalization can match examinees to tasks relevant for them and, consequently, enhance engagement, motivation to perform, and the activation of prior knowledge, then performance should theoretically be higher relative to assessment methods that do not have similar mediating effects (González et al., 2005; Hefflin, 2002; Lee, 1998). In a quasi-experimental study, Wang et al. (2025) compared the scores of Black and non-Black

students on a reading test form centered around the Harlem Renaissance and on forms focusing on other topics deemed less relevant to Black students. Wang et al. found Black students to be more engaged in the Harlem Renaissance form and to show smaller score differences relative to their non-Black peers. An experimental study could be conducted to evaluate this research question more rigorously, as well as the mediating effects. The study could include random assignment of examinees who self-identify with selected demographic groups to ETS Testlet forms constructed to be more or less relevant to those groups (e.g., because reading passages were written by famous group members or because the passages describe key aspects of the group's history). Engagement, motivation to perform, relevant prior knowledge, and Testlet performance would be measured and compared for each form-by-group condition. The expectation would be that demographic groups taking forms that were constructed to be more relevant would be more engaged, motivated, and would have prior knowledge activated and thus would score higher than members of the same demographic group taking a less-relevant form (i.e., differential boost; Sireci et al., 2005). It is also possible to evaluate the interaction hypothesis—that is, not only should administration of relevant tasks produce higher scores for diverse students relative to traditional measures, but a smaller increase (or no increase) should also be observed for White students (because the forms were constructed to be relevant to other groups). Cognitive lab methods could also be used to supplement the evaluation of prior-knowledge activation in student samples from selected group-by-form conditions.

Study 4

Research Question. Does an increase in performance on a personalized assessment that maximizes task relevance lead to higher levels of self-confidence and sense of efficacy?

Summary. Bandura (1977) cited performance accomplishments as a major source of information contributing to self-efficacy (see, e.g., Caprara et al., 2011, for evidence of a virtuous cycle between self-efficacy beliefs and course grades). As such, an increase in test performance, on top of heightened engagement and motivation, should positively impact self-confidence and sense of efficacy. This study might be best conducted as part of Principle 1, Study 3 (though the added burden to examinees of additional measures could make that idea

infeasible). Should Study 3 produce positive results, an alternative might be to follow it with a simplified replication. This follow-up would randomly assign examinees from minoritized and nonminoritized groups to two ETS Testlet forms, with one form intended to be more relevant to the minoritized group than the other form. Following their receipt of performance results, measures of efficacy and confidence would be given to both groups. Analyses would compare confidence and efficacy, as well as performance, across the four group-by-form combinations (i.e., minoritized group with a relevant form, minoritized group with a less-relevant form, nonminoritized group with a minoritized relevant form, nonminoritized group with a minoritized less-relevant form).

Study 5

Research Question. Do examinees and other stakeholders perceive personalized assessment that maximizes task relevance as more or less fair than current assessment methods?

Summary. Fairness has been defined in many ways (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). One way to conceive of it is as a perception of appropriate treatment given the treatment afforded to others. This study will ask members of various stakeholder groups (e.g., educators, parents, examinees) to make online judgments of the fairness of personalized assessments that try to match examinees to tasks relevant to their cultural identity, interests, prior knowledge, or background. Several different approaches to that matching will be described. One such approach is to select or generate items using AI methods that are intended to align with known examinee characteristics. A second approach is to let examinees bring their characteristics to the response by presenting items suitably open in their response requirements. A third approach is to combine these two methodologies, as in asking the examinee's input as to what aspects of the problem and response mode might be personalized and in what manner. In each case, stakeholder judgments will be made relative to traditional assessment methods that do not attempt such a match. Respondents will be asked to judge fairness in a Likert format and write a short explanation for each comparison.

Principle 2: Allow for Multiple Forms of Representation and Expression in Problem Stimuli and in Responses

Thoughtfully incorporating multiple forms of representation, and permitting alternate modes of expression, should cause students to show what they know and can do more than would be apparent under the typically limited means of expression and representation provided on standardized tests (Ketterlin-Geller, 2005; Sireci et al., 2005), thereby increasing performance and the perceptions of all stakeholders that testing is fair. In addition, this incorporation should increase student engagement to the degree that desired forms and modes are available for examinees to choose. Finally, problems that aid students in making deep-structure connections among representational forms and expressive modes should enhance the chances for subsequent transfer of learning, as well as improved test performance (Carpenter, 2012).

Study 1

Research Question. Does choice of mode of expression or form of representation positively impact performance?

Summary. Allowing alternate modes of expression (spoken, written, sign) and forms of representation (braille, large print, language versions) has been a long-standing practice for examinees with disabilities and for English learners, falling within the general rubric of Universal Design for Assessment (Ketterlin-Geller, 2005). The foundational notion is to give students multiple ways to access content and multiple ways to respond. Extending that notion more broadly might involve permitting a greater variety of modes and representational forms to be used in, for example, responses to problems calling for the demonstration of content knowledge. That is, if the measurement target is basic knowledge of circulatory system function, there is no strong rationale for requiring a textual response. That knowledge might be demonstrated more readily for some students via a drag-and-drop diagram, bulleted list, or oral recitation. This affordance might be of particular benefit to those who are English learners. It may also benefit anyone whose written expression is relatively limited, whether due to typing, handwriting, or verbal dysfluencies. This study would use content-based questions consistent with the state learning standards for participants' grade levels.

A possible design for this study would be to have participants answer questions presented in randomly assigned pairs, one question answered in their choice of modes of expression or representational form (e.g., bulleted list, brief essay, drag-and-drop diagram) and the other question in one of the response types they did not select, with presentation order counterbalanced across pairs. To control for variation in difficulty, attain adequate sample size for each question–mode combination, and increase the power of statistical tests, an algorithm would keep track of how often each pair of item–mode combinations appeared (where a pair includes the two combinations an examinee answered). The algorithm would then assign the first item and the item–mode combination for the second item that has appeared less often. Analyses of the resulting data would focus on identifying whether performance was better for the preferred mode or form.

Study 2

Research Question. Does choice of mode of expression or form of representation positively impact examinees' engagement?

Summary. Giving agency to examinees to choose the mode of expression or form of representation should theoretically increase engagement and motivation (cf. self-determination theory; Ryan & Deci, 2000, 2023), thus decreasing cognitively disengaged response behavior. Cognitively disengaged response behavior, or rapid response behavior, can be defined as an unrealistically fast response. Such responses suggest that the examinee did not complete the (meta)cognitive processing required to seriously consider the problem (Arslan & Finn, 2023; Finn, 2015; Wise, 2017). An experimental study could compare the proportion of disengaged responses (i.e., noneffortful responses), as measured by time on task, across conditions in which examinees are and are not given the choice of mode or form. This study could be conducted in combination with Principle 2, Study 1 or done independently.

Study 3

Research Question. Do examinees and other stakeholders perceive personalized assessment that permits choice of mode of expression or form of representation as fairer than current assessment methods?

Summary. Because of its simplicity and use of the same target population (i.e., stakeholders), this study could possibly be combined with that of Principle 1, Study 5. Participants would be shown exemplars of a question that offers choice of mode of expression and representational form. They would then be asked to rate its fairness relative to the same question offering no choice of mode or form.

Study 4

Research Question. Would providing multiple forms of representation or modes of expression in assessment encourage teachers to instruct how to choose among representations and modes? Would it encourage teaching for transfer (i.e., instruction to encourage competency application across modes or forms)?

Summary. Assessment drives instruction in that teachers and students tend to focus their efforts more on the content and formats represented on upcoming tests than on unrepresented content and formats (Ainsworth, 2018). Thus, we would expect that wider incorporation in tests of multiple forms of representation and modes of expression would impact teaching and learning practice. This research question might be investigated by interviewing a sample of teachers virtually to find out how their teaching practice might change given the appearance on their state assessment, or on college admissions tests, of choice among representations and modes. A structured interview protocol might be employed to probe as to whether and how teachers adjust their instruction to help students make beneficial choices and/or attempt to develop in their students a more ready facility to move among modes and representations in problem solving. A control condition might involve presenting to the same or a randomly parallel sample of teachers test problems that do not include as wide an incorporation of modes and representations. Qualitative comparisons between the two conditions should suggest the impact of this assessment manipulation on teaching. Although this study would indicate only what teachers *say* they would do, the outcome should suggest whether a follow-up investigation collecting data from actual classroom practice might be worthwhile.

Study 5

Research Question. Does providing multiple forms of representation or modes of expression on the assessment facilitate the transfer of knowledge or skill?

Summary. If the presence of multiple modes and forms on the assessment influences teaching and learning behavior, then students would be expected to engage in classroom practice aimed at recognizing how the same problem and solution can be represented in those multiple forms. Under such conditions, students should perform better on assessments that offer this variety than students not given such practice. To test this hypothesis, we might give training tasks to one group of students that include multiple representational forms and expressive modes, where intentional deep-structure connections are promoted across the modes or forms. For example, the task might be to choose from among five problems in different representations (e.g., verbal, graphical, symbolic) the pair having the same deep structure. In the control condition, all five problems would take a single representational form (i.e., only verbal, only graphical, only symbolic), the task again being to match the two problems with the same deep structure. In both conditions, students would then be given novel problems of the type administered in both conditions to see if and where transfer occurred (Bransford & Schwartz, 1999). Students who can efficiently transfer their learning would be expected to retrieve and apply appropriate deep structural knowledge and skills quickly and competently toward solving the novel problems (Schwartz et al., 2005). Students' performance and strategies used in solving the transfer problems could be compared across conditions to provide evidence of whether such inclusion facilitates transfer compared to conditions in which only a single type of representation or expressive mode is utilized.

Principle 3: Promote Instruction for Deeper Learning Through Assessment Design

Promoting deeper learning through assessment design should cause teachers unfamiliar with approaches to such instruction to begin to incorporate these approaches in their practice. In conjunction with teachers giving greater attention to deeper learning, modeling such learning in the assessment should cause students to increase meta-cognitive self-regulatory behavior, including monitoring their performance against quality standards and internalizing the processes employed by proficient domain performers (Frederiksen, 1984; Resnick &

Resnick, 1990; Shepard, 2021). These changes in student and teacher behavior should lead to greater learning.

Study 1

Research Question. Do assessment designs that facilitate deeper learning via performance tasks and real-world resources lead to changes in teachers' instructional practice in ways that enhance students' test performance and support their learning?

Summary. As Bennett (2023) has suggested, designing assessments to foster deeper learning may involve the inclusion of real-world performance tasks and supportive resources that students can consult in the process of solving significant domain-relevant problems. It is important to investigate the extent to which teacher and student interactions with such performance tasks influence teaching and learning practice within the classroom. In partnership with one or more schools or districts, this issue could be investigated by incorporating multiple performance tasks, exemplified by scenario-based assessments developed under the *CBAL*™ research initiative (Bennett et al., 2018) or ETS Testlets (O'Dwyer et al., 2023), into classroom practice over several instructional units. This study could involve a combination of teacher surveys, interviews, and classroom observations in addition to collection of data from student interactions with the performance tasks. The aim would be to obtain evidence of the degree to which teachers incorporated the deeper learning techniques embedded in the assessments into their current and subsequent instructional units and assessment practices.

Study 2

Research Question. What are the effects of different types of feedback on students' deeper learning practices?

Summary. Personalized assessments can give students feedback intended to affect deeper learning practices (Hattie, 2009; Maier & Klotz, 2022). That feedback can be provided during (and after) the assessment experience. Potential study designs may involve comparing the performance of students receiving various types of feedback (e.g., conditions that prime attention specifically to problems' deep structure or to criteria for quality performance vs. more general feedback not associated with deeper learning practices). Test performance,

engagement, and motivation levels can be compared across such conditions. In addition, such a study could examine the degree to which feedback that supports deeper learning fosters students' meta-cognitive and self-regulatory behaviors. These behaviors could be assessed in various ways, including post assessment teacher report, student self-report, or investigation of actions taken within the assessment as captured in log files.

Principle 4: Adapt the Assessment to Student Characteristics

Adapting to personal characteristics should cause stakeholders to feel that the assessment is fairer because it aligns better with student interests, cultural identity, background, and prior knowledge than does a traditional test. Adaptation should also cause higher levels of motivation and engagement with the test (e.g., Bernacki & Walkington, 2018; Walkington, 2013; Walkington & Bernacki, 2020), thereby increasing examination performance. Given appropriate guidance, allowing choice should enhance competency in taking effective agency, which should, in turn, positively affect learning (Brod et al., 2023; National Academies of Sciences, Engineering, and Medicine, 2018; Patall, 2013; Patall et al., 2017; Shepard, 2021). To the extent that agency encourages examinees to explore cultural identity and share their explorations, those identities should be reinforced and sustained. The greater the degree of adaptation to the aforementioned personal characteristics is, the larger should be the salutary effects, especially for students from traditionally underserved groups.

Study 1

Research Question. Do examinees perceive personalized assessments that allow for extensive agency as more or less reinforcing of identity relative to traditional methods?

Summary. Approaches to personalization for equity are based in part on a premise derived from work on teaching diverse students, which is the idea of reinforcing and sustaining cultural identity (Paris, 2012; Paris & Alim, 2014). The most obvious examples of assessments that might do so are those allowing extensive agency, that is, permitting the examinee to bring to the assessment whatever aspects of their identity, background, prior knowledge, and interests they choose. Examples of operational assessments that offer extensive agency include the AP Research (College Board, 2023a) and AP Art and Design Portfolio (College Board, 2023b)

examinations. In those examinations, the affordance of agency allows examinees to employ cultural identity as a vehicle for demonstrating competency in the focal construct (Bennett, 2023). In AP Art and Design, for example, artworks that reflect examinee identity are readily identifiable (Escoffery et al., 2025). In AP Research, project topics that target the interests of specific demographic groups are also presumably choices based at least in part on cultural identity (e.g., one student's project was titled "The Link Between Asian American Portrayal in the Media and Euro-American Historical Views of Asians"). This study will explore examinee perceptions with respect to the extent to which personalized approaches that privilege examinee agency might work for or against promoting identity as compared with conventional methods. The research question might be addressed by asking participants to self-identify as members of minoritized or nonminoritized groups. Next, they could be shown a collection of prompts, each of which would afford a different degree of agency, and asked to rate the extent to which each might allow them to respond in a way that would meaningfully engage their identity. Additional questions might probe whether they preferred prompts that allowed such affordance and, if so, whether they thought such affordance might help reinforce and sustain their identity.

Study 2

Research Question. Does giving agency to examinees to personalize the context of the task increase motivation, engagement, and performance?

Summary. Previous studies on context personalization in mathematics instruction have shown positive effects on learners' performance and motivation (e.g., Bernacki & Walkington, 2018; Walkington, 2013). This study would explore those effects in assessment. With the help of generative AI, on-the-fly context personalization can give examinees agency by allowing them to personalize the task setting *during* the assessment based on their interests and cultural identity, holding construct-related task demands constant (Arslan, 2024; Arslan et al., 2024). For example, in mathematics, story problems already include a predefined context. This context can be relevant to some examinees' experience but not to the experience of others. Giving examinees agency to personalize that context during the assessment should increase motivation and engagement (see self-determination theory; Ryan & Deci, 2000, 2023), thus

allowing them to better show what they know and can do. This claim could be evaluated by comparing the examinee's motivation, engagement, and performance across two conditions. In the “agency” experimental condition, participants would be allowed to personalize the problem context before answering. Examinees in the “no agency” control condition would be assigned personalized problems based on their background characteristics.

Study 3

Research Question. Does guiding students in taking agency help them make better choices among tasks?

Summary. Prior studies that permitted examinees to choose among assessment prompts have documented that some examinees fail to make beneficial choices (Powers & Bennett, 1999); that is, they choose problems on which they score lower than they would have scored on other, unchosen problems. However, it ought to be the case that gently guiding examinees with respect to evaluating options will result in better choices. Such guidance might be provided in various ways, and this study would test the effectiveness of only one of those ways. The study hypothesis might be examined experimentally by asking examinees to rate short-text-response prompts on the extent to which each prompt calls upon relevant prior knowledge they possess and on how interesting the prompt is to them, as well as to explain each rating briefly. Examinees would then be asked to answer the prompt of their choice, after which they would be asked to respond to one other randomly assigned prompt. Participants in the control condition would be asked to rate each prompt on variables that should be less relevant to making good choices (e.g., number of words, number of punctuation marks), then choose a prompt, respond to it, and answer a second, randomly assigned prompt. Analyses would compare performance between chosen and randomly assigned prompts within and across conditions.

Study 4

Research Question. What is the impact of specific conditions around giving examinees choice on examinee motivation, engagement, and performance?

Summary. While giving examinees agency through choice is expected to increase their motivation, engagement, and performance generally, the specific conditions under which these benefits might occur remain unclear, as do the populations with whom those conditions might interact. Conditions of interest might include the assessment context (e.g., large-scale, formative classroom), whether and when to receive feedback, feedback type, and whether and how often to offer choice to personalize the contexts of the tasks.

Controlled studies can systematically explore the impact of such specific conditions on motivation, engagement, and performance. For instance, researchers have studied the effects of different types of feedback (e.g., knowledge of results, knowledge of correct response, elaborated feedback, and answer-until-correct; Fong et al., 2019; Mertens et al., 2022; Shute, 2008; Van der Kleij et al., 2015). However, these studies usually assign students to one of the feedback conditions and compare it with a no-feedback condition. A study could be conducted by giving examinees a choice of the types of feedback compared to conditions in which examinees are randomly assigned to a feedback type.

Under Principle 4, Study 2, we proposed an experiment to investigate the use of generative AI for on-the-fly context personalization, which gives examinees agency to change the task context during the assessment based on their interests as embedded in their cultural identities. The expected positive effects of such context personalization might have a diminishing return as a function of the number of tasks examinees are asked to personalize. Moreover, some examinees may have no interest in personalizing the context at all. A study could be conducted to systematically vary the number of tasks examinees need to personalize, including no such option (i.e., taking standard tasks without personalization). Such a study would examine the effects of this variation on students' motivation, engagement, and performance.

Such investigations could also incorporate measures of students' perceived agency. The goal would be to investigate hypotheses around the extent to which providing agency of different kinds and degrees fosters a more general sense of agentic competency, which is predicted to improve test performance and learning, redounding to increases in sense of self-efficacy.

Study 5

Research Question. Do examinees and other stakeholders perceive personalized assessment that adapts to examinee characteristics as fairer than current assessment methods?

Summary. Similar to Principle 1, Study 5 and Principle 2, Study 3, this investigation would evaluate stakeholder perceptions of assessment approaches that offer choice. Participants would be shown question exemplars that give the examinee a choice of adaptations to individual characteristics. Adaptations would be selected that are likely to be salient to a relevant sample of students given the role of the stakeholder (e.g., for teachers of English learners, availability of a second-language glossary, dual-language presentation of item content). Stakeholders would then be asked to rate the fairness of that choice relative to the same question without choice of adaptation to the pertinent individual characteristics.

Principle 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

Characterizing 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 should cause examinees, teachers, the public, and policymakers to interpret, communicate about, and act on assessment results more carefully than is currently the case. More careful interpretation means recognizing that, absent other evidence, results are bound to task types, conditions, and contexts like those employed in the assessment—selections that developers should have made on a defensible basis and justified. Understanding results as an interaction should cause students to know better how task features, conditions, and contexts affect their performance. Similarly, that knowledge should lead teachers and students to experiment with modifications of these factors that facilitate learning and improve test performance.

Study 1

Research Question. Do teachers, parents, students, school administrators, and policymakers make more appropriate interpretations when assessment results are characterized as an interaction?

Summary. The premise underlying this study is that characterizing performance as an interaction will cause stakeholders to make more bounded interpretations of results and, therefore, to be more likely to take justifiable actions (Kay et al., 2020). To test this claim, an experimental study could be conducted in which results are reported as an interaction to one group of stakeholders (e.g., school administrators) and reported more generally to a randomly parallel group. Each group would then be given a small number of selected-response questions as to the interpretation and use of the results. As an example intended for adult participants, the result given to the control group might be as follows:

> The 2011 NAEP writing assessment was given on computer and called upon students to respond to three writing purposes: to persuade, explain, or convey experience. Females achieved a mean scale score of 160 and males scored 140. The difference between the female and male groups was statistically significant. Which of the following interpretations is most justifiable?
>
> (a) U.S. eighth-grade females could be considered to be better writers than males when composing on-demand online essays to persuade, explain, or convey experience;
>
> (b) U.S. eighth-grade females could be considered to be significantly better writers than U.S. eighth-grade males across writing genres;
>
> (c) U.S. eighth-grade females could be considered to have received significantly better writing instruction than did U.S. eighth-grade males.

This question would be followed with a prompt asking the participant to explain the reason for selecting their interpretation.

For the *experimental* group, the prompt would be reworded to emphasize the interaction:

> On the 2011 NAEP writing assessment, when composing online essays on demand to persuade, explain, or convey experience, females achieved a mean scale score of 160 and males scored 140. The difference between the female and male groups was statistically significant. Which of the following interpretations seems most justifiable?

This prompt would be followed by the same response options as given to the control group, along with a request to explain their reasoning.

For each group, a follow-up question might focus on appropriate action:

Based on the results, which of the following is the most appropriate action to recommend?

(a) Fund research studies to identify the causes for the difference in performance between U.S. eighth-grade females and males;

(b) Fund professional development for teachers in how to teach writing to U.S. eighth-grade males more effectively;

(c) Fund instructional initiatives directed at more effectively motivating U.S. eighth-grade males to write.

If the experimental group more often chooses option (a), coupled with reasoning that explicitly mentions the bounded nature of the assessment results, the hypothesis would be supported that interactional framing leads to more careful interpretations. Implementation of this study would need to be careful to avoid potential spillover effects between conditions due to participants interacting, if it was run in the same physical setting.

Study 2

Research Question. Can examinees be primed to understand assessment results as an interaction so that they know better how task features, conditions, and contexts might affect their performance?

Summary. Task features, conditions, and contexts make a difference. That fact is illustrated by the so-called person-by-task interaction (Linn & Burton, 1994; Shavelson et al., 1993), which occurs when two tasks of similar average difficulty function such that one task is easy for Examinee A but hard for Examinee B and the other task is hard for Examinee A and easy for Examinee B. Person-by-task interaction is the underlying basis for personalized assessment—that is, attempting to match task demands to examinee background, interest, identity, and other relevant characteristics to see what an examinee knows and is able to do

under optimal circumstances. Calling examinees' attention to task features, conditions of administration, and administration contexts, and considering those factors with respect to their own resources, may help examinees understand their task performance and how it might vary given such factors. That understanding becomes critical in criterion situations in which task demands are *not* personalized but come as they are, potentially creating suboptimal situations. In these situations, reflective examinees can attempt to adjust the demands to better fit their resources, expand their resources to meet the task demands more effectively, or both.

In this structured interview study, secondary school students will be given questions that encourage them to analyze task and situational demands and reflect on the resources they bring to meet those demands (i.e., to understand better the nature of performance as an interaction between specific demands and the resources they might marshal to meet them). The literature on examinee choice offers no examples of reflection questions that might serve as suitable models for a structured interview protocol. Consequently, the questions and procedure that follow are meant only to suggest possibilities for motivating thought on the part of interested investigators.

Each reflection question would be answered in response to writing prompts. The reflection questions, some of which would branch, might be like the following:

- Here are three brief writing prompts, each focused on a different writing purpose. Have you responded to similar prompts in school before? If so, which ones? Which of the three purposes do you feel most prepared to write about? Why? What could you do to better prepare yourself to write for the other purposes?
- Here are two writing prompts, each focused on the writing purpose you just selected but dealing with different topics. Which of the two topics do you feel more prepared to write about? Why? What aspects of your background or interests might be relevant to the topic that you could use in responding? What could you do to better prepare yourself to write about the *other* topic?
- With respect to the second of the two writing prompts just presented, if you could write your response on paper or computer, which mode do you feel would allow you to demonstrate your writing skills better? Why? If you had to write your

response in the other mode, what might you do that would make the writing task easier?

- This writing prompt could be taken on computer with 20-minute or 40-minute time limits. Which limit do you think would allow you to better demonstrate your writing skills? Why? If you had to take the task under the other time limit, how might you prepare in advance to ensure your best performance? What might you do in the writing session itself to ensure your best performance?

After responding to the preceding questions, participants will be asked to respond to a question similar to that used in Principle 5, Study 1, about the interpretation of results from a writing assessment:

> A timed writing assessment was given on computer. The assessment called upon students to respond to three writing purposes: to persuade, explain, or convey experience. Females achieved a considerably higher score than males. Which of the following interpretations of the test results is most justifiable?
>
> (a) Females are better writers than males when composing timed online essays to persuade, explain, or convey experience;
>
> (b) Females are better writers than males regardless of writing genre, so they perform better on timed writing tests;
>
> (c) Females received better writing instruction than males, so they perform better on timed writing tests.
>
> What reasoning led you to choose that response? Is there any connection between the questions you were asked earlier about writing prompts and your response to this question? If so, explain that connection.

These questions are intended to probe the extent to which the participant understands assessment results as an interaction and whether the reflective questions played a role in helping to advance that understanding.

Study 3

Research Question. Does understanding assessment results as an interaction lead teachers and students to be more willing to experiment with modifications of these factors to facilitate learning and improve test performance?

Summary. If teachers and students comprehend the underlying premise of personalization, they should realize that assessment results represent an interaction (Zapata-Rivera et al., 2007). That interaction is among the resources the student brings to the assessment, the tasks, the context, and the conditions of administration. Optimizing the match between resources internal to the student and external factors should cause better performance. In academic and workplace settings, individuals will sometimes encounter situations in which the factors are somewhat malleable (e.g., as in choice of task or conditions), thereby allowing a degree of optimization. At other times, individuals may face challenges that must be taken as given. Because personalized assessment attempts to optimize performance, results might not represent how a student would perform in situations that pose less favorable matches. As such, teachers and students would do best to regularly experiment with modifications to tasks, conditions, and contexts to facilitate learning and improve performance regardless of the match. Such improvement could come from students being taught to make wiser selections in choice situations to better match factors to their resources or, alternatively, making efforts to increase their competency in dealing with challenging task features, conditions, and contexts. Both directions suggest that teachers and students engage in reflective practice.

The adult-learning literature has examples of questionnaires designed to gauge such practice (e.g., Gustafsson et al., 2021; Larrivee, 2008; Priddis & Rogers, 2017). These questionnaires, some of which were developed for teachers, typically pose statements to be rated by respondents on a Likert scale. In general, such scales pose questions not well suited to the purposes and context of this study (e.g., “When reflecting with others about my work I become aware of things I had not previously considered”). Thus, in what follows we offer illustrative questions more directly aligned to the study purpose and context. This set is

intended only to provide the interested researcher a starting point in conceptualizing how the problem might best be approached.

In this structured interview study, secondary school teachers would be presented with a hypothetical situation and questioned about it in a manner similar to the following:

> When you assess a student, where do you get the questions that you use? Do you use the same questions for each student in the class? If you modify the questions, why do you do that? What types of modifications do you make?
>
> Think of a particular student in your class. What is special about that student? Does that student have interest in a particular area? Is that student especially knowledgeable about some topic(s) outside of school? Does that student come from a family that is known to have notable cultural interests or practices?
>
> Imagine that you are interested in learning more about how well this student can write under optimal conditions. Here is a writing prompt. How could you modify this task to probe what that particular student knows and can do? What specific aspects of the task would you change?
>
> What if you changed the context so that it better matched the student's background, interests, or cultural identity? If you are not familiar with the culture from which the student comes, is there a colleague who might help you with that modification? How might you find out more about the student's cultural background, identity, and experiences?
>
> What if you changed the timing so that it allowed for more initial planning?
>
> Now imagine that you wanted to know how well this student can write under less optimal conditions, conditions more like the ones they might encounter in postsecondary education or the workplace.
>
> Consider the same writing prompt. How might you change the prompt so that it doesn't match that student's background, interests, or identity? What if you changed the context so that it came from a culture with which the student was not familiar? What if

you changed the topic to one in which you knew the student was not interested? How do you think the student would perform?

What might you do to get the student to make his or her own task modifications so that the student begins to experiment with how those factors affect performance? Do you think that understanding those effects might lead the student to make wiser selections when choice among tasks was offered, thereby allowing the possibility of a better match between task factors and student resources? Do you think that if this student understood better the relationship between resources and external task factors, the student might be motivated to try to increase their competency in dealing with challenging task features, conditions, and contexts? In what other ways might you encourage that student to try to develop this competency?

When you assess a student in the future, where will you get the questions? Will you use the same questions for each student in the class? Why? If you modify the questions, why will you do that? What types of modifications do you think you might make?

An associated structured interview study could also be conducted with a small number of students along the same lines as described earlier, the goal being to increase awareness of the nature of assessment results as interactions.

The analysis will be qualitative, concentrating on the extent to which teachers and students show (a) awareness of assessment results as an interaction and (b) willingness to experiment with modifying task features, conditions, and contexts in the future.

Study 4

Research Question. How does a personalized learning system built on an integrated learner model help students and teachers understand performance as an interaction among learner characteristics and task features?

Summary. Personalized assessment based on multiple learner characteristics (personal, social, cultural, linguistic) requires the use of an integrated learner model (Lehman et al., 2024; Sparks et al. 2024). That model accounts for how the interaction of learner and assessment characteristics affects (a) students' assessment (or task) experiences and (b) how students

perform. An *open* learner model can make that complex interaction visible by presenting it in a way that is easily understandable to teachers and students (Bull & Kay, 2016). Such a presentation, appropriately guided, might help them better understand the interactive nature of performance results.

This study would involve two phases. In the first phase, a particular assessment scenario would be used to ground the task of determining what learner characteristics are most relevant for personalization. Using that scenario, an integrated learner model would be codesigned with teachers and students separately. For example, the assessment scenario could be a formative, classroom-based assessment of geometry. Teachers and students would be asked to identify what characteristics within each of four categories (personal, social, cultural, linguistic) might be most relevant for personalization and/or contextualization of performance results. The process would then continue to determine perceptions as to how the identified characteristics interact to influence assessment experience and performance. Next, teachers and students would brainstorm ways in which they might want to engage with feedback reports, focusing on the interaction between learner characteristics, assessment features, testing experience, and test performance. This process may require multiple codesign sessions.

The second phase of the study would begin with members of the research team developing multiple mock-ups based on the codesign work. Mock-ups would be divided into two sets, one for teachers and one for students, with the members of each set containing different paths to engagement with feedback reports (i.e., user journeys). Teachers and students (separately) would then engage in structured interviews to walk through each user journey to (a) identify the one with which they most connect (or to explain the way in which they would prefer to interact with the open learner model if none of the user journeys fit), (b) provide insights into their interpretations of the information in the open learner model (see other studies under Principle 5 for example questions), and (c) identify the appropriate methods for presenting test performance as an interaction between learner characteristics and test features.

Study 5

Research Question. Does an AI-based interactive score report help students and teachers understand performance as an interaction among learner characteristics and task features?

Summary. Research in the areas of interactive score reporting and open learner models has provided insights on design principles to support understanding and appropriate use of assessment results (Bull & Kay, 2016; Kannan & Zapata-Rivera, 2022). This study will present teachers and students with an AI-enhanced report intended to facilitate their understanding of the interactive nature of assessment results, as well as other aspects of performance. The AI-enhanced report will provide personalized interpretive text and engage in text-based conversations with users about the assessment results that appear on the screen. Questions will be posed to users about how performance might have changed if modifications had been made to specific task features, conditions, and administrative contexts. For example, a teacher might be asked how student performance would be expected to change if (a) the problem contexts were modified so that they better matched the student's background, interests, or cultural identity; (b) the contexts were removed entirely; (c) the test timing was relaxed; (d) the student was asked to orally give and discuss responses; or (e) the problems were presented to small groups for solution instead of to the student alone. Semi-structured interviews will be employed to gather information about the extent to which the interactive report fostered understanding of assessment results, particularly regarding their interactional character.

Conclusion

This research memorandum described an investigatory agenda for personalized assessments that are sensitive to the social, cultural, and other relevant characteristics of individuals and the contexts from which they come. The agenda targeted two major purposes. The first purpose was to guide the theoretical and empirical research and development needed to create assessments that differentially adjust to the needs of individuals. The second purpose was to suggest a means for judging the efficacy of those assessments.

The agenda was intended to be a general one, applying across the full range of assessment uses—school accountability, national and international assessment, admissions,

certification and licensure, and instructional planning—with customizations to those purposes as appropriate. The goal of such assessments is to measure more validly the competencies of all individuals, but especially those from minoritized groups (see Randall et al., 2022).

The studies were framed around a theory of socioculturally responsive assessment. That theory, which was derived from multiple literatures, represents one way in which personalized assessment in the service of equity might be conceptualized. For purposes of the agenda, the theory was used to motivate the selection of research questions, the questions being tests of the theory's propositions. Using the theory's design principles, personalized assessments can be built and evaluated, with efficacy judged on the extent to which the designs theoretically align with the principles and on the degree to which the empirical data support the propositions derived from those principles.

Several topics key to personalized assessment were not addressed by this agenda. One such topic is measurement methodology, including how to link scores from disparate assessments so that they are comparable enough for intended purposes, how to evaluate the quality of those scores, and how to compute their precision. Some progress on these topics has been made by Sinharay and Johnson (2024), Sinharay, Bennett, et al. (2025), and Sinharay, Johnson, et al. (2025). A second key topic is computational modeling, covering such issues as how to represent in an integrated learner model an individual's characteristics and interactions with the assessment, as well as how to employ that model in real time to adjust content, format, response modality, and other assessment conditions. Each of these topics deserves its own research and evaluation agenda comparable in scope to the current one.

The current agenda should be considered a starting point for conceptualizing a broad, comprehensive research and evaluation program on personalizing assessment in the service of equity. That research and evaluation program should span the substantive studies described here, as well as the measurement methodology and computational modeling topics expected to be described in future documents. As studies are conducted and results assembled, new questions will be raised, novel or amended theoretical frameworks will be proposed, and further studies will be conceived to push forward the possibilities of personalized assessment in the service of equity.

References

Ainsworth, S. (2018). Multiple representations and multimedia learning. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 96–105). Routledge. https://doi.org/10.4324/9781315617572

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

Arslan, B. (2024, May 27–28). *Personalized, adaptive, and inclusive digital assessment and learning environments* [Conference presentation]. E-ADAPT Conference, Potsdam, Germany. https://osf.io/82p5f/?view_only=cba3f410bc1e462fb086e3361ffed0bc

Arslan, B., & Finn, B. (2023). The effects of personalized nudges on cognitively disengaged student behavior in low-stakes assessments. *Journal of Intelligence*, *11*(11), Article 204. https://doi.org/10.3390/jintelligence11110204

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence*, *7*, Article 1460651. https://doi.org/10.3389/frai.2024.1460651

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

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. (2024). Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice*, *43*(4), 119–125. https://doi.org/10.1111/emip.12652

Bennett, R. E. (2025). Rethinking equity and assessment through opportunity to learn. *Assessment in Education: Principles, Policy, and Practice*, *32*(1), 1–28. https://doi.org/10.1080/0969594X.2025.2462549

Bennett, R. E., Zwick, R., & van Rijn, P. (2018). Innovation in K–12 assessment: A review of CBAL research. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment:*

Development, modeling, and scoring from an interdisciplinary perspective (pp. 197–247). Information Age. https://doi.org/10.1108/978-1-68123-931-620251010

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology*, *110*(6), 864–881. https://doi.org/10.1037/edu0000250

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*(1), 61–100. https://doi.org/10.3102/0091732X024001061

Brod, G., Kucirkova, N., Shepherd, J., Jolles, D., & Molenaar, I. (2023). Agency in educational technology: Interdisciplinary perspectives and implications for learning design. *Educational Psychology Review*, *35*, Article 25. https://doi.org/10.1007/s10648-023-09749-x

Bull, S., & Kay, J. (2016). SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, *26*, 293–331. https://doi.org/10.1007/s40593-015-0090-8

Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, *81*(1), 78–96. https://doi.org/10.1348/2044-8279.002004

Carpenter, S. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

College Board. (2023a). *AP Research: Course and exam description*. https://apcentral.collegeboard.org/courses/ap-research

College Board. (2023b). *AP 2-D Art and Design, 3-D Art and Design, Drawing: Course and exam description*. https://apcentral.collegeboard.org/courses/ap-3-d-art-and-design

College Board. (2023c). *AP US History: Course and exam description*. https://apcentral.collegeboard.org/courses/ap-united-states-history

College Board. (2023d). *AP Computer Science Principles: Course and exam description*. https://apcentral.collegeboard.org/courses/ap-computer-science-principles/course

Ebe, A. (2010). Culturally relevant texts and reading assessment for English language learners. *Reading Horizons: A Journal of Literacy and Language Arts*, *50*(3), 193–210. Retrieved from https://scholarworks.wmich.edu/reading_horizons/vol50/iss3/5

Ebe, A. (2025). Examining the relationship between the cultural relevance of text and reading proficiency: Using a cultural relevance rubric in reading assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 168–179). Routledge. https://doi.org/10.4324/9781003435105

Escoffery, D. S., Fletcher, K. E., & Stone-Dahany, R. (2025). "A search for my voice": Socioculturally responsive assessment in AP Art and Design. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 262–282). Routledge. https://doi.org/10.4324/9781003435105

Evans, C. M. (2023). Applying a culturally responsive pedagogical framework to design and evaluate classroom performance-based assessments in Hawai'i. *Applied Measurement in Education*, *36*(3), 269–285. https://doi.org/10.1080/08957347.2023.2214655

Finn, B. (2015). *Measuring motivation in low-stakes assessments* (Research Report No. RR-15-19). ETS. https://doi.org/10.1002/ets2.12067

Fong, C. J., Patall, E. A., Vasquez, A. C., & Stautberg, S. (2019). A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review*, *31*(1), 121–162. https://doi.org/10.1007/s10648-018-9446-6

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*(3), 193–202. https://psycnet.apa.org/doi/10.1037/0003-066X.39.3.193

Gay, G. (2018). Culturally responsive teaching: Theory, research, and practice (3rd ed.). Teachers College Press. https://doi.org/10.31046/wabashcenter.v1i3.1798

González, N., Moll, L. C., & Amanti, C. (Eds.). (2005). Funds of knowledge: Theorizing practices in households, communities, and classrooms. Lawrence Erlbaum Associates Publishers.

Gustafsson, S., Engström, Å., Lindgren, B. M., & Gabrielsson, S. (2021). Reflective capacity in nurses in specialist education: Swedish translation and psychometric evaluation of the Reflective Capacity Scale of the Reflective Practice Questionnaire. *Nursing Open*, *8*(2), 546–552. https://doi.org/10.1002/nop2.659

Hattie, J. A. C. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hefflin, B. R. (2002). Learning to develop culturally relevant pedagogy: A lesson about cornrowed lives. *Urban Review*, *34*, 231–250. https://doi.org/10.1023/A:1020603323594

Kannan, P., & Zapata-Rivera, D. (2022). Facilitating the use of data from multiple sources for formative learning in the context of digital assessments: Informing the design and development of learning analytic dashboards. *Frontiers in Education*, *7*, Article 15. https://doi.org/10.3389/feduc.2022.913594

Kay, J., Zapata-Rivera, D., & Conati, C. (2020). The GIFT of scrutable learner models: Why and how. In A. M. Sinatra, A. C. Graesser, X. Hu, B. Goldberg, & A. J. Hampton (Eds.), *Design recommendations for intelligent tutoring systems* (Vol. 8, pp. 25–40). U.S. Army Combat Capabilities Development Command—Soldier Center. Available from https://www.researchgate.net/profile/Jeanine-Defalco-2/publication/

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), 1–22. Available from https://files.eric.ed.gov/fulltext/EJ848519.pdf

Ladson-Billings, G. (2021). *Culturally relevant pedagogy: Asking a different question.* Teachers College Press. Available from https://www.tcpress.com/culturally-relevant-pedagogy-9780807765920

Larrivee, B. (2008). Development of a tool to assess teachers' level of reflective practice. *Reflective Practice*, *9*(3), 341–360. https://doi.org/10.1080/14623940802207451

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education*, *67*(3), 268–279. http://www.jstor.org/stable/2668195

Lehman, B., Sparks, J.R., Zapata-Rivera, D., Steinberg, J., & Forsyth, C. (2024). A culturally enhanced framework of caring assessments for diverse learners. *Practical Assessment, Research, & Evaluation, 29*(1), 9. https://doi.org/10.7275/pare.2102

Lin, L., Lin, X., Zhang, X., & Ginns, P. (2024). The personalized learning by interest effect on interest, cognitive load, retention, and transfer: A meta-analysis. *Educational Psychology Review*, *36*, Article 88. https://doi.org/10.1007/s10648-024-09933-7

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5–8. https://doi.org/10.1111/j.1745-3992.1994.tb00778.x

Maier, U., & Klotz, C. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, *3*, Article 100080. https://doi.org/10.1016/j.caeai.2022.100080

Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. *British Journal of Educational Technology*, *52*(5), 1935–1964. https://doi.org/10.1111/bjet.13116

Malda, M., van de Vijver, F. J. R., & Temane, Q. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, *38*(6), 582–595. https://doi.org/10.1016/j.intell.2010.07.004

Mertens, U., Finn, B., & Lindner, M. A. (2022). Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. *Journal of Educational Psychology*, *114*(8), 1743–1772. https://doi.org/10.1037/edu0000764

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & Vendlinski, R. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation*, *19*(2–3), 121–140. https://doi.org/10.1080/13803611.2013.767614

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press. https://doi.org/10.17226/24783

National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academies Press. https://doi.org/10.17226/9853

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8*. National Academies Press. https://doi.org/10.17226/11625

O'Dwyer, E., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Applied Measurement in Education*, *36*(3), 286–303. https://doi.org/10.1080/08957347.2023.2214652

Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational Researcher*, *41*(3), 93–97. https://doi.org/10.3102/0013189X12441244

Paris, D., & Alim, S. (2014). What are we seeking to sustain through culturally sustaining pedagogy? A loving critique forward. *Harvard Educational Review*, *84*(1), 85–100. https://doi.org/10.17763/haer.84.1.982l873k2ht16m77

Patall, E. A. (2013). Constructing motivation through choice, interest, and interestingness. *Journal of Educational Psychology*, *105*(2), 522–534. https://doi.org/10.1037/a0030307

Patall, E. A., Vasquez, A. C., Steingut, R. R., Trimble, S. S., & Pituch, K. A. (2017). Supporting and thwarting autonomy in the high school science classroom. *Cognition and Instruction*, *35*(4), 337–362. https://doi.org/10.1080/07370008.2017.1358722

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

Priddis, L., & Rogers, S. L. (2017). Development of the reflective practice questionnaire: Preliminary findings. *Reflective Practice*, *19*(1), 89–104. https://doi.org/10.1080/14623943.2017.1379384

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting White supremacy in assessment: Towards a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Resnick, L. B., & Resnick, D. P. (1990). Tests as standards of achievement in schools. In *Proceedings of the 1989 ETS Invitational Conference: The uses of standardized tests in American education* (pp. 63–80). ETS. Available from ED335421.pdf

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryan, R. M., & Deci, E. L. (2023). Self-determination theory. In F. Maggio (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 6229–6235). Springer. https://doi.org/10.1007/978-3-031-17299-1_2630

Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–51). Information Age.

Sexton, U., & Solano-Flores, G. (2002, April 1–5). *Cultural validity in assessment: A cross-cultural study on the interpretation of mathematics and science test items* [Paper presentation]. American Educational Research Association annual meeting, New Orleans, LA, United States.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator*, *45*(3), 28–37. https://www.aft.org/ae/fall2021/shepard

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sinharay, S., Bennett, R. E., Kane, M., & Sparks, J. R. (2025). Validation for personalized assessments: A threats-to-validity approach. *Journal of Educational Measurement*, *62*(2), 282–310. https://doi.org/10.1111/jedm.12434

Sinharay, S., & Johnson, M. S. (2024). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. *Journal of Educational Measurement*, *61*(1), 5–46. https://doi.org/10.1111/jedm.12381

Sinharay, S., Johnson, M. S., Bennett, R. E., Lopez, R. M., Sparks, J. R., & Pillarisetti, S. (2025). Investigating elements of culturally responsive assessments in the context of the National Assessment of Educational Progress: An initial exploration. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.70008

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*(4), 457–490. https://doi.org/10.3102/00346543075004457

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, *28*(2), 9–18. https://doi.org/10.1111/j.1745-3992.2009.00143.x

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

Sparks, J. R., Lehman, B., & Zapata-Rivera, D. (2024). Caring assessments: Challenges and opportunities. *Frontiers in Education*, *9*, Article 1216481. https://doi.org/10.3389/feduc.2024.1216481

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932–945. https://doi.org/10.1037/a0031882

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of*

Research on Technology in Education, *52*(3), 235–252. https://doi.org/10.1080/15391523.2020.1747757

Wang, Z., Sparks, J., Walker, M., O'Reilly, T., & Bruce, K. (2025). Group differences across scenario-based reading assessments: Examining the effects of culturally relevant test content. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 369–398). Routledge. https://doi.org/10.4324/9781003435105

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Zapata-Rivera, D., Hansen, E., Shute, V. J., Underwood, J. S., & Bauer, M. (2007). Evidence-based approach to interacting with open student models. *International Journal of Artificial Intelligence in Education*, *17*(3), 273–303. https://doi.org/10.3233/IRG-2007-17(3)04

Zheng, L., Long, M., Zhong, L., & Gyasi, J. F. (2022). The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: A meta-analysis. *Education and Information Technologies*, *27*(8), 11807–11830. https://doi.org/10.1007/s10639-022-11092-7

Notes

[1] *Minoritized group* refers to any social group treated as subordinate to the dominant social group.

[2] Even so, such groups may be used in specific studies for testing theoretical propositions and design principles.

[3] Text describing the propositions is based on Bennett (2023) and is used by permission.

[4] This memorandum, in most cases, uses *examinee* and *student* interchangeably in that most of the assessment purposes envisioned are educational in nature. Where those terms are not used interchangeably, the intent is to imply either a testing context (examinee) or a formative classroom assessment context (student).

www.ets.org