Research Report

Reflections on Equity-Centered Design

ETS RR–20-22

María Elena Oliveri
Jessica Nastal
David Slomp

December 2020

ETS Research Report 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.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Reflections on Equity-Centered Design

María Elena Oliveri[1], Jessica Nastal[2], & David Slomp[3]

1 Independent Consultant
2 Prairie State College, Chicago Heights, IL
3 University of Lethbridge, Coledale, AB Canada

This report discusses frameworks and assessment development approaches to consider fairness, opportunity to learn, and consequences of test use in the design and use of assessments administered to diverse populations. Examples include the integrated design and appraisal framework and the sociocognitively based evidence-centered design approach. The report also provides an overview of approaches that have been used before to increase the utility, relevance, and authenticity (verisimilitude) of assessments to inform various decisions made from the use of assessments administered to diverse populations. These considerations are important as the test-taker populations engaging in assessment situations become increasingly more diverse at the national and international levels.

Keywords Fairness; evidence-centered design; integrated design and appraisal framework; sociocognitive; relevant; authentic

doi:10.1002/ets2.12307

The coordinated session on equity-centered design at the annual meeting of National Council on Measurement in Education (NCME) in April 2019 addressed questions related to inequity and unfairness in assessment, the challenges found with respect to assessment, and especially the assessment of students of color. The session also discussed potential ways in which assessments may privilege particular groups and the audience's perspectives for what assessment can be. In this report, we reflect on these questions as we describe challenges associated with the use of assessments with diverse populations. We also discuss approaches to consider fairness and consequences of test use to enhance the assessments' utility, relevance, and purpose.

We build on previous research conducted on test fairness and social justice from the fields of writing studies, given the long-standing consideration of the test takers' diverse needs in these fields. Perhaps more so than assessments of other constructs, writing assessments tend to draw on student backgrounds such as when students are asked to express their opinion on issues. Nonetheless, the arguments and notions raised in this report also apply to assessments of other constructs.

On the topic of writing studies, we refer to questions raised in *Writing Assessment, Social Justice, and the Advancement of Opportunity* in relation to the advancement of opportunity to learn, the need to complement large-scale assessments with classroom assessments, and the importance of highlighting ecological, rhetorical, and sociocognitive frameworks to interpret the writing process in a meaningful manner (Poe et al., 2018). Poe et al. (2018) also discussed the importance of advancement of opportunity, the identification of opportunity structures, and actionable outcomes in educational contexts. They also highlighted the needed connection between assessment and instruction (see also Gee, 2008; Moss et al., 2008).

A focus on expanding opportunity to learn and the achievement of social justice in writing assessment suggests that we move beyond traditional test fairness review processes, which often involve examining items by investigating differential item functioning (DIF) posttest administration. Ercikan and Oliveri (2013) and Oliveri et al. (2014) identified shortcomings related to DIF analyses when tests are administered to heterogeneous populations leading to underdetecting DIF. These findings suggest revising traditional test fairness review processes from a focus on analyzing DIF postadministration to considering test fairness from initial test development stages. *The ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations* (International Test Commission, 2018) also suggests an emphasis on considering potential sources of bias during test development when assessments are administered to diverse populations. The need to reconceptualize and expand fairness perspectives in assessment is also elaborated in the frameworks described next.

Corresponding author: M. Oliveri, E-mail: oliveri.m@live.com

Frameworks Considering Fairness in Assessment for Diverse Populations

Considerations of diverse test-taker populations' needs from the early stages of test design are discussed in guidelines and frameworks (Banks et al., 2018; Elliot, 2016; International Test Commission, 2018; Mislevy, 2018; Poe et al., 2019; Slomp, 2016). The foundation has been long established in writing studies. In examining teaching and learning in basic writing classes, Bossone (1967), for example, urged instructors to understand their students' background and "language characteristics" (p. 90): "through records and reports, as well as interviews, the instructor should learn as much as [they] can about the student—biologically, psychologically, and environmentally" (p. 90) to offer instruction that facilitates each student's learning. Stenglass's (1997) longitudinal study dramatized how students' material conditions; racial, ethnic, and cultural background; disability status; and personal lives are intertwined with their academic performance, and Rose's work (e.g., 2012) shows the dialectical tensions between students' personal and academic contexts.

Slomp (2012) observed that when the focus of assessment shifts from measuring the products test takers produce to making inferences about the development of specific abilities (such as writing), a bioecological lens (Bronfenbrenner & Morris, 2006) is needed to help make sense of the interpersonal, intrapersonal, and ecological factors that either support or inhibit that development. More recently, Gallagher (2014) observed "writing assessment must be conducted by those who know something about writing and who live most directly with the consequences of assessments" (pp. 487–488). And, as White et al. (2015) explained, assessment must necessarily "be—at the very least—sensitive to context" (p. 166) because it requires us to "attend to the impact of what we do" (p. 166); and even more significantly, "vitality is to be found in context" (p. 143).

A central goal of the frameworks is to guide the assessment development efforts in support of meaningfully contextualizing test questions for students of diverse backgrounds and the associated interpretation of diverse students' responses to avoid penalizing a student's response due to sources of construct-irrelevant variance (CIV). The minimization of CIV is important to allow for valid and fair score-based interpretations when assessing diverse populations. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) defined CIV as "variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation" (p. 217). The challenge, however, of determining which factors are extraneous and which factors are construct relevant should not be underestimated. Foundational to equity-centered assessment design, therefore, are (a) robust, detailed construct maps emphasizing span across cognitive, intrapersonal, interpersonal, and neurological domains (National Research Council, 2012; White et al., 2015) (b) an understanding of the populations assessed in terms of complex factors of intersectionality (Carbado & Harris, 2019; Crenshaw, 1991) and (c) a clear understanding of the matrix of bioecological factors that can impact test-taker performance (Zumbo et al., 2015). Taken together, this information can assist assessment designers and users to understand what they are measuring, what bioecological factors may impact performance (including which of these factors are construct relevant and which are construct irrelevant), and how various test-taker populations perform on the assessment. Relying on an ethical standpoint (Mislevy & Elliot, 2020) to advance social justice (Elliot, 2016) and opportunity to learn for all students (Moss et al., 2008), the sociocognitive approach depends on detailed prior understandings of the constructs and assessed populations.

Given the importance of minimizing CIV and the consideration of various factors impacting assessment use and score-based interpretations, a number of models for assessment design and appraisal, grounded in a sociocognitive approach have been developed (Mislevy, 2018; Slomp, 2016). Each of these models contributes and informs an equity-centered design approach. An example is the integrated design and appraisal framework (IDAF; Slomp, 2016), which we highlight in this paper because of its concomitant emphasis on fairness and consequences of test use. Later, we also highlight a resonant model: sociocognitive design (Mislevy, 2018).

Integrated Design and Appraisal Framework

IDAF draws on Kane's (2013, 2016) model of validation, Mislevy et al.'s (2003) evidence-centered design approach, and White et al.'s (2015) design for assessment framework. Assessment design based on the IDAF places the principle of fairness in the foreground of the design process. It emphasizes the need to consider the intended and unintended consequences of assessment design choices at each of an assessment's design, implementation, and use phases. The objectives are to maximize positive intended outcomes of assessments and minimize negative ones. As shown in Table 1, assessment design based on the IDAF model proceeds in six phases, appraisal of the assessment program follows the same six phases.

Table 1 Phases of Assessment Design, Development, and Appraisal

Phase	Design and Development	Appraisal
1	Define aims, principles of fairness, and context	Review aims, principles of fairness, and context
2	Identify elements foundational in the assessment design	Analyze the assessment's core construct and content domains
3	Develop an assessment program	Critically examine assessment design with respect to construct representation
4	Develop scoring system	Critically examine scoring procedures with respect to impact on construct representation
5	Develop a plan for analyzing assessment results	Critically examine the evidence related to the chain of inferences supporting assessment use; particular attention is paid to disparate impact with respect to populations of interest
6	Develop a plan for analyzing assessment consequences	Collect and review evidence related to the intended, unintended, positive, and negative outcomes stemming from the assessment's design and use

In Phase 1 of the design process, the assessment purpose and the populations affected by the assessment are defined. This phase includes identifying the information needs the assessment is designed to fill, the inferences to be made based on the data the assessment generates, and the populations of interest the assessment will affect. In Phase 2, the constructs of interest and content domains are identified and mapped. Stability of the construct is examined particularly with respect to the populations of interest identified in Phase 1. In Phase 3, the construct and content domains inform the blueprint for the assessment, and items are designed. Once designed, items are then critically reviewed to identify possible sources of CIV they introduce into the assessment. Assessment blueprints are reviewed with an eye to construct underrepresentation. In Phase 4, scoring criteria and scoring procedures are developed. Scoring criteria are mapped to the construct and content maps created in Phase 2. Scoring procedures are examined to minimize the chances of potentially introducing construct underrepresentation or irrelevant variance into the assessment program. Criteria and procedures are also examined to help ensure they do not introduce CIV, especially as this relates to populations of interest. In Phase 5, intended scoring, generalization, and extrapolation inferences are reviewed. The arguments, warrants, and potential causal relationships needed to support each inference are defined and tested. A sampling plan is developed to ensure each population of interest identified in Phase 1 is represented in sufficient quantity to allow for a meaningful analysis of their performance in relation to other populations. In Phase 6, intended outcomes of the assessment program are reviewed and a plan for collecting evidence to determine if those outcomes are being realized is developed. At this stage, a plan for uncovering and responding to unintended outcomes for individuals, populations, and educational systems is also developed.

During the appraisal cycle, the six phases are revisited with a critical eye. Aims are reviewed and populations of interest are identified, construct and content maps are critically reviewed, and sampling plans for populations of interest and for test items are critically reviewed. If disaggregation of results demonstrates differences in performance by populations of interest evidence based on test content, test-taker response processes and internal structure need to be collected and reviewed to determine whether the assessment is measuring the same construct across populations. The chain of inferences also needs to be reviewed to help ensure that disaggregated performance data do not undermine the scoring, generalization, and extrapolation inferences. If differences in performance across populations is not a result of problems with the instrument, a research plan for collecting evidence related to opportunity to learn for each disparate population must be developed and enacted. This research plan should provide direction for how best to structure opportunity for these populations to help raise performance for diverse groups.

As a proof of concept case study of IDAF, an analysis of the student learning assessment (SLA) Program, a Grade 3 formative literacy and numeracy assessment program, developed by the government of the province of Alberta (Canada) was completed in 2017 (Slomp et al., 2017). The study generated 17 lessons and 11 recommendations for how the SLA program could be improved. The absence of a systematic focus on fairness in the design and implementation of the SLA program became apparent during this process, leading the evaluation team to recommend:

> Alberta Education should commit to a program of research that focuses on the issue of fairness as it relates to issues of validity and reliability. This program of research should examine performance by subgroup in both literacy and

> numeracy, (a) to understand why models are substantially weaker for some performance groups, (b) to identify ways to strengthen the assessment models through item refinement, and (c) to draw lessons from high performing groups that could be applied to enhancing the performance of other subgroups. (Slomp et al., 2017, p. 9)

While the assessment findings are beyond the scope of this paper, results suggested that the IDAF framework served to advance equity-centered design in a principled way.

A Sociocognitive Perspective to Assessment Development

Another relevant and complementary framework for equity-centered design is the sociocognitive approach to the evidence-centered design framework. As Mislevy (2018) suggested, ignoring diversity in examinees' thinking, reasoning, and responses to test items can lead to alternative explanations for outcomes for subgroups within populations. Mislevy argued that the awareness of the social context should permeate all aspects of the assessment process, including construct conceptualization, its evidentiary rules, analyses, and findings so that the heterogeneity associated with different groups' thinking, response processing, and responding can be captured more accurately. The sociocognitive approach calls developers to attend to key elements of task design, construct representation, and the type of resources and knowledge culturally and linguistically diverse populations might bring to the assessment (Mislevy, 2018; O'Sullivan & Weir, 2011; Weir, 2005). Central goals of a sociocognitive approach to test design are to guide decisions (e.g., which tasks to include so that the context used in the task is familiar to diverse populations, how to frame the test language, visual representations, and scoring rules) so that appropriate score-based inferences are obtained for the multiple test-taker populations (Oliveri et al., 2019).

Understood in resonance with the design elements of IDAF and their emphasis on consequence, Mislevy's sociocognitive framework provides a necessary psychometric underpinning—especially in terms of the mathematical expression of the construct model and their expression as authentic or real-world models are approximated by considering sociocognitive perspectives (Mislevy & Elliot, 2020). To elaborate, the sociocognitive approach is operationalized in the IDAF framework, and that framework is articulated psychometrically taking into account sociocognitive considerations.

Considerations for Making Assessments More Useful for Diverse Populations

Central to addressing the coordinated session's goals, beyond the description of existing models and frameworks designed to consider equity-centered design, is to identify key principles that can help render assessments useful and relevant for diverse populations. Thus, the next section aims to respond to the question raised in the coordinated session regarding what the assessments at issue can be. To address this issue, we consider the following questions: Do the assessments (a) assess relevant and useful skills? (b) provide useful information that supports teaching and learning, and (c) yield information of relevance for the various test-taker populations and subgroups?

Relevance: Do the Tests Assess Relevant and Useful Skills?

Central to developing and using educational assessments is to ask whether the tests assess relevant and useful skills, which involves developing assessments with robust construct representation that focus on identifying valued habits of mind and skills; that is, skills that are needed in current and future-looking contexts (e.g., Levy, 2010). Analyses of relevance may be informed by national and international multidisciplinary groups of experts (e.g., economists, educators, and policymakers) to develop not only relevant frameworks that suggest the types of skills needed for success today, but also analyze how these targeted skills may be building blocks for skill sets needed for the future for students' academic, professional, personal, and civic development.

Moreover, it is important to consider the way the skills will be assessed and characterized. For instance, certain skills can be better characterized and taught when a context is provided for their learning and when they are connected to real-world problems. Otherwise, students may find it difficult to see the connection between what is taught and real-world applications. Unfortunately, more traditional forms of assessment and the associated teaching-to-the-test phenomena has led to the assessment of less relevant skills in ways that are distant from real-world applications (Gorin & Mislevy, 2013).

Thus, framework development must be grounded in systematic reviews of the empirical research literature that investigates what knowledge, skills, and dispositions lead to expertise with respect to the construct of interest. Otherwise, negative unintended consequences may follow. Students that graduate with inappropriate preparation or without mastery of the applied skills and knowledge expected in the workforce, and employers may struggle to fill positions with candidates who have the relevant competencies needed for today's and tomorrow's workplace (Casner-Lotto & Barrington, 2006; Hart Research Associates, 2015; Slomp et al., 2014). These negative consequences may also extend to students missing out on additional benefits higher education brings, as Rose (2012) suggested, "from improved health and health literacy, to reduced crime rates, to enhanced quality of life for the students and their families" (p. 47).

An associated question about skill relevance and its characterization asks about relevance for whom? Are all test-taker populations able to interpret the items meaningfully? Would all test takers have been involved with similar opportunities to interact with the skills and the technologies used to deliver the skills? For example, Murphy (2007) reported on an instance where children in a rural setting were silent when an assessor asked them to identify a picture of a sheep because they "didn't know whether he wanted to know if the sheep was a pure Chevoit or a Crossbreed" (p. 238). In this case, the children's silence may have been misconstrued as lack of understanding or knowledge of the test item when in fact it shows deeper understanding than what is traditionally expected for the item.

As the test-taker population diversifies, our field is presented with new challenges requiring us to develop/adapt test interpretation frameworks, new demands to structure opportunities to learn, and increased opportunities to innovate with respect to assessment design and development (Ercikan & Oliveri, 2016). These steps may help ensure the tasks remain relevant, so we obtain accurate performance estimates of complex skills. This innovation is more technically demanding than the creation of discrete tasks. It will also require multiple observations of performances across multiple assessment tasks. It is possible that such innovations would make our community and the associated products more useful and valuable, because they would be viewed as enablers rather than barriers to educating students that are of diverse educational, linguistic, and cultural backgrounds (Oliveri & Mislevy, 2019).

Moreover, innovations may come through the development of contextualized tasks that enable learners to acquire deeper problem-solving capabilities based on real-world representations and scenarios—a movement gaining ground in writing placement and developmental education reform, as well as economics and moral philosophy (see Nussbaum, 2011). This focus is particularly important for underserved students, as they may have not received the types of experiential and problem-solving education, which can help better prepare students with the needed skills (e.g., communication, problem solving) for an evolving economy (Mehta, 2014).

Thus, to enhance student preparation, ensure workplace readiness, advance opportunity to learn, and investigate interaction with the test items and technology used to deliver them, formal methods are required. For instance, we suggest that when assessments are administered to diverse populations, skill characterizations and contextualization consider a sociocognitive perspective to assessment development operationalized in IDAF.

To illustrate, Oliveri et al. (2019) discussed design choices associated with the development of collaborative problem-solving assessments for multiple populations, suggesting that developers consider approaches to depicting characters in scenario-based tasks, including how the characters interact with each other and the type of language used in character interactions. In the scenario described in the article based on a scenario-based task for Arabic, American, and Canadian test takers in the oil and gas extraction industry, the authors suggested considering whether women and men are to be depicted in the scenario as working together or separately and the extent to which to use formal or informal language in chats given that Arabic is a diglossic language. Thus, "failing to consider the inclusion or exclusion of cultural or linguistic differences in the test-taker populations during task design may lead to inaccurate score-based inferences or a misjudgment of individuals' abilities" (p. 272). These considerations are important within a sociocognitively extended evidence-centered design framework.

To explain the force of such modeling, we now elaborate on implications related to opportunity to learn and the development of assessments that better integrate teaching and learning. The assessments may thus aim to provide information that supports teaching and learning by providing more meaningful feedback to the stakeholders (Gorin & Mislevy, 2013).

Utility: Do Assessments Provide Useful Information to Inform Instruction?

Beyond making the assessments more relevant, we propose a focus on expanding the amount and type of information assessments provide to stakeholders. Examples of such practices may include the use of formative assessments, the use

of data that involves assessing skills that build on each other across assessment administrations, and providing feedback about the test takers' progress in meaningful ways (Shepard, 2006; Wiliam, 2006). For users of scores and assessment data (e.g., test takers, teachers, tutors), formative assessments can have a stronger relationship with learning than summative assessments and can be helpful although they may require increased testing time and multiple test administrations.

To provide further specificity regarding what those types of assessments would look like, we provide two examples: (a) the Berkeley Evaluation and Assessment Research (BEAR) assessment system and (b) the Cognitively Based Assessment of, for, and as Learning (*CBAL*®) system. Shared features of these systems include assessments that are embedded in classroom instruction and lead to enhanced learning opportunities for teachers and students. Black and Wiliam (1998) suggested that effective teaching practices involve providing professional development to teachers to support effective formative assessment practices and providing teachers with an understanding of diverse learning models, assessment strategies, and instructional decision-making. Thus, in the BEAR assessment system, teachers are involved in the development of the assessments, are provided with a set of tools for assessing student learning over time, and are given professional development of the use of these tools to provide feedback to inform student learning.

Along these lines, the CBAL research initiative has explored in-depth how summative and formative assessment can be designed from the ground up to support learning (see Bennett, 2010). One aspect of this research has focused on how an explicit model of literacy skill develops, built on the idea of key practices (such as argumentation, research, or literary analysis). Through this research, learning progressions were developed in collaboration with teachers. The objectives are to (a) create a variety of tasks to illustrate many of the features that support effective instruction and (b) inform the development of contextualized classroom scenario-based tasks. These complex performance tasks are broken down into scaffolded sequences of subtasks that teach strategies modeling skilled literacy practices and promoting deeper processing through the use of graphic organizers and planning and organizational tools, as well as rubrics, checklists, and other support and evaluation tools (Deane et al., 2008, 2015; Hayes, 2006; Prior, 2006). The CBAL formative materials support multiple modes of interaction (individual, group, collaborative, and whole-class work) with some classroom implementations that require students to co-construct a shared response or review one another's response and provide feedback. The tasks also are intended to provide actionable information to teachers, focused on the skills and strategies students need to acquire to reach higher levels of performance on CBAL English language arts (Deane et al., 2008, 2015; Hayes, 2006; Prior, 2006) learning progressions.

Diversity: Do Assessments Yield Information of Relevance for Subgroups?

Central to developing assessments that yield information that is relevant to participating subgroups is to analyze the extent to which the test-taker groups are heterogeneous (see Elliot, 2016). Based on expanded domain models, analyses may include examining possible differences in the learning strategies of the test takers, differences in the background knowledge, life experiences of the multiple populations, collaborative ability, and the varied ways that differently abled students approach complex problems. Moreover, psychometrically, researchers may wish to use analyses such as cluster analyses, factor analyses, and item response theory mixture models to discover whether there may also be instructionally/cognitively relevant subgroups in the data, such as strategy use, or curricular emphases that are not known from available covariates (Mislevy, 2018). The complementary use of these models also can add insight regarding differences across subgroups.

As the heterogeneity of the test-taking population increases, the assumptions, decisions, and interventions informed by test scores may not necessarily apply to all members of a subgroup. If such misalignments occur at a systemic level and impact test-taker groups, the data may be used inappropriately or the wrong interpretations may be made leading to unintended negative consequences (e.g., poor resource allocations, limited educational access, and interventions that are off-target) for some subgroups. These considerations are particularly relevant to special populations or subgroups to whom policies are intended to apply, including English language learners (Kopriva & Sexton, 1999), linguistic minorities (Oliveri & von Davier, 2015), and students with learning disabilities (Laitusis et al., 2011). Disaggregating results of assessments to determine patterns of performance in test-taker populations can help to identify issues of bias or opportunity to learn in the assessment contexts. The analysis of construct remodeling data collected from special populations of test takers can help to decide if differences in performance between groups arise from issues of bias (see for example, Fox & Cheng, 2007) or from issues related to opportunity to learn. If CIV is identified as the problem, test designers and users can identify its potential sources and make modifications as needed to the test items. If opportunity to learn is identified as

the problem, practitioners are required to "structure opportunity and thereby advance opportunity to learn" (Elliot, 2016, section1.1, para. 1)—an invitation to innovate. Prior analysis of bioecological factors impacting subgroups can help to identify barriers to development that must be addressed.

Concluding Thoughts

In closing, we note that as the test-taker populations continue to become more diverse and the skills that are useful and needed continue to expand, it is incumbent upon us to innovate methods and practices of assessment that provide for greater opportunity to learn for all students. With a shift toward equity-centered design, we hope to create more meaningful assessment experiences that draw on students' rich backgrounds and provide them with ample opportunity to demonstrate their knowledge. Equity-centered design foregrounds the essential human aspect of our work. It can lead us to understand more about why communities of students perform better or worse than others and how they come to understand what they are asked to do. Some of that understanding can come from direct interactions with learners and professionals in various fields of study and research through multidisciplinary communities of practice that involve assessment developers, psychologists, and instructors who are vested in addressing fundamental questions such as: How can we improve teaching and learning? How can we honor each learner's unique background? How can we improve our test development practices to make assessments more meaningful and useful?

Addressing these questions is not only a national concern but also an international one. As we continue to explore and advance equity-centered design models, researchers will continue to collaborate—as the authors have in this paper—on lessons learned to gain a richer understanding of the strategies that worked in their own countries. Needed internationally are efforts to design assessments that consider the needs of diverse populations from both national and international perspectives (Oliveri & Wendler, 2020).

Acknowledgments

The first author thanks Jennifer Randall for the invitation to give a presentation at her coordinated session on equity-centered design (Oliveri, 2019). The current report, which was written while the first author was on staff at Educational Testing Service, builds on discussion of that presentation.

References

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

Banks, W. P., Burns, M. S., Caswell, N. I., Cream, R., Dougherty, T. R., Elliot, N., Gomes, M., Hammond, J. W., Harms, K. L., Inoue, A. B., Lederman, J., Molloy, S., Moreland, C., Nulton, K. S., Peckham, I., Poe, M., Sassi, K. J., Toth, C., & Warwick, N. (2018). The braid of writing assessment, social justice, and the advancement of opportunity: Eighteen assertions on writing assessment with commentary. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 379–425). The WAC Clearinghouse, University Press of Colorado. https://wac.colostate.edu/docs/books/assessment/braid.pdf

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*(2–3), 70–91. https://doi.org/10.1080/15366367.2010.508686

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, *5*, 1–57.

Bossone, R. M. (1967). Remedial English in junior colleges: An unresolved problem. *College Composition and Communication*, *18*(2), 88–93.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology: Vol.1. Theoretical models of human development* (6th ed., pp. 793–828). Wiley. https://doi.org/10.1002/9780470147658.chpsy0114

Carbado, D. W., & Harris, C. I. (2019). Intersectionality at 30: Mapping the margins of anti-essentialism, intersectionality, and dominance theory. *Harvard Law Review*, *132*(8), 2193–2239.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. Partnership for 21st Century Skills.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (Research Report No. RR-08-55). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02141.x

Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Yi., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Foley, C. (2015). *Key practices in the English language arts: Linking learning theory, assessment and instructio*n (Research Report No. RR-15-17). Educational Testing Service. https://doi.org/10.1002/ets2.12063

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, 9(1). http://journalofwritingassessment.org/article.php?article=98

Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 69–86). Emerald Publishing.

Ercikan, K., & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education*, *29*(4), 310–318. https://doi.org/10.1080/08957347.2016.1209210

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario secondary school literacy test by first and second language test-takers. *Assessment in Education*, *14*(1), 9–26. https://doi.org/10.1080/09695940701272773

Gallagher, C. (2014). Review essay: All writing assessment is local. *College Composition and Communication*, *65*(3), 486–505.

Gee, J. P. (2008). A sociocultural perspective on opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 76–108). Cambridge University Press. https://doi.org/10.1017/CBO9780511802157.006

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf

Hart Research Associates. (2015). *Falling short? College learning and career success: Selected findings from online surveys of employers and college students conducted on behalf of the Association of American Colleges & Universities*. https://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf

Hayes, J. (2006). New directions in writing theory. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 28–40). The Guilford Press.

International Test Commission. (2018). *ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations*. https://www.intestcom.org/page/31

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64–80). Routledge.

Kopriva, R., & Sexton, U. M. (1999). *A guide to scoring LEP student responses to open-ended science items*. Council of Chief State School Officers.

Laitusis, C. C., Buzick, H. M., Cook, L. L., & Stone, E. (2011). Adaptive testing options for accountability assessments. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margins: Challenges, strategies, and techniques* (pp. 291–310). Information Age Publishing. https://www.infoagepub.com/products/Assessing-Students-in-the-Margin

Levy, F. (2010). *How technology changes demands for human skills* (OECD Education Working Paper No. 45). http://www.oecd.org/edu/skills-beyond-school/45052661.pdf

Mehta, J. (2014, June 20). *Deeper learning has a race problem*. Education Week. http://blogs.edweek.org/edweek/learning_deeply/2014/06/deeper_learning_has_a_race_problem.html

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

Mislevy, R. J, & Elliot. N. (2020). Ethics, psychometrics, and writing assessment: A conceptual model. In J. Duffy & L. Agnew (Eds.), *After Plato: Rhetoric, ethics, and the teaching of writing* (pp. 143–162). Utah State University Press.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective*, 1(*1*), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802157

Murphy, S. (2007). At last: Culture and consequences: The canaries in the coal mine. *Research in the Teaching of English*, *42*(2), 228–244. https://www.jstor.org/stable/40171726

National Research Council. (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. *The* National Academies Press. https://doi.org/10.17226/13398

Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Harvard University Press.

Oliveri, M. E. (2019, April 4 – 8). *Equity-centered design*. Paper presented at the annual National Council on Measurement in Education Conference, Toronto, ON, Canada.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286 – 300. https://doi.org/10.1080/08957347.2014.944305

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, *19*(3), 270 – 300. http://doi.org/10.1080/15305058.2018.1543308

Oliveri, M. E. & Mislevy, R. (2019). Introduction to "challenges and opportunities in the design of 'next-generation assessments of 21st century skills'" special issue. *International Journal of Testing*, *19*, 97 – 102. https://doi.org/10.1080/15305058.2019.1608551

Oliveri, M. E., & von Davier, A. A. (2015). Psychometrics in support of a valid assessment of linguistic minorities: Implications for the test and sampling designs. *International Journal of Testing*, *16*(3), 220 – 239. https://doi.org/10.1080/15305058.2015.1069743

Oliveri, M. E., & Wendler, C. (Eds.). (2020). *Higher education admissions practices: An international perspective*. Cambridge University Press. https://doi.org/10.1017/9781108559607

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), Language testing: Theories and practices *(pp. 13 – 32)*. Palgrave Macmillan.

Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*. The WAC Clearinghouse, University Press of Colorado. https://wac.colostate.edu/docs/books/assessment/justice.pdf

Poe, M., Nastal, J., & Elliot, N. (2019). Reflection. An admitted student is a qualified student: A roadmap for writing placement in the two-year college. *Journal of Writing Assessment*, *12*(1). http://journalofwritingassessment.org/article.php?article=140

Prior, P. (2006). A sociocultural theory of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 54 – 66). Guilford Press.

Rose, M. (2012). *Back to school: Why everyone deserves a second chance at education*. The New Press.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623 – 646). American Council on Education, Praeger.

Slomp, D. (2016). An integrated design and appraisal framework for ethical writing assessment. *Journal of Writing Assessment*, *9*(1). http://www.journalofwritingassessment.org/article.php?article=91

Slomp, D., Elliot, N., Marynowski, R., & Rudniy, A. (2017). *Alberta's student learning assessment program: An integrated evaluation. Executive Summary*. https://education.alberta.ca/media/3615918/sla-research-executive-summary.pdf

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing*, *17*(2), 81 – 91. https://doi.org/10.1016/j.asw.2012.02.001

Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English*, *48*(3), 276 – 302. https://www.jstor.org/stable/24398680

Stenglass, M. S. (1997). *Time to know them: A longitudinal study of writing and learning at the college level*. Erlbaum.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave MacMillan.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Utah State University Press.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, *11*, 283 – 289.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136 – 151. https://doi.org/10.1080/15434303.2014.972559

Suggested citation:

Oliveri, M. E., Nastal, J., & Slomp, D. (2020). *Reflections on equity-centered design* (Research Report No. RR-20-22). Educational Testing Service. https://doi.org/10.1002/ets2.12307

Action Editor: Donald Powers

Reviewers: Michael Kane and Robert Mislevy