



**Wisconsin Center for  
Education Research**  
SCHOOL OF EDUCATION  
UNIVERSITY OF WISCONSIN-MADISON

# **Developing Culturally Responsive Assessments of Teaching: Implications of Bob Mislevy**

**WCER Working Paper No. 2026-2**  
**August 2026**

---

**Courtney Bell and Drew Gitomer**

Wisconsin Center for Education Research  
University of Wisconsin–Madison  
[courtney.bell@wisc.edu](mailto:courtney.bell@wisc.edu)

*Keywords:* culturally responsive assessment, teacher evaluation, evidence-centered design, validity, culturally responsive teaching, validity argument, teaching quality, educational equity

*Suggested citation:* Bell, C. & Gitomer, D. (2026). *Developing culturally responsive assessment of teaching: Implications of Bob Mislevy* (WCER Working Paper No. 2026-2). University of Wisconsin–Madison, Wisconsin Center for Education Research.

© 2026 by Courtney Bell and Drew Gitomer. All rights reserved. Any opinions, findings, or conclusions expressed in this paper are those of the author and do not necessarily reflect the views of the Universities of Wisconsin, WCER, or cooperating institutions. Readers may make verbatim copies of this document for noncommercial purposes by any means, provided that the above copyright notice appears on all copies. WCER working papers are available at <https://wcer.wisc.edu/publications/working-papers>

# **Developing Culturally Responsive Assessments of Teaching: Implications of Bob Mislevy**

**Courtney Bell and Drew Gitomer**

## **ABSTRACT**

Bob Mislevy spent decades developing theory, methods, and tools that positioned assessment as an instance of the more general problem of using evidentiary argument to make inferences in the presence of uncertainty. He argued that skills and understandings are situated within social contexts and therefore, the purpose of an assessment must guide any assessment use. This paper applies Mislevy's foundational questions—what are we assessing, for what purpose, and how should we build the assessment—to the domain of teaching. We define teaching as a situated complex adaptive system unfolding across three domains (classroom organization, social-emotional support, instructional support and activation) and argue that culturally responsive assessments must evaluate teaching, not teachers. We propose that such assessments should serve two purposes—supporting teacher learning and supporting teacher flourishing—and should be designed and validated through participatory, iterative, context-aligned processes that engage stakeholders throughout, not only during construct definition. Drawing on Kane's validity-argument framework and Mislevy's evidence-centered design, we specify how interpretive and validity arguments for teaching assessments must attend to task design, scoring warrants, contextual information about teachers, and alternative explanations. We conclude that culturally responsive assessment and Mislevy's sociocognitive approach are mutually reinforcing resources for advancing more just assessments of teaching.

# Developing Culturally Responsive Assessments of Teaching: Implications of Bob Mislevy

Courtney Bell and Drew Gitomer

Bob Mislevy spent the last several decades developing theory, methods, designs and tools to enrich the breadth and quality of assessment. He was not simply trying to re-engineer and improve traditional assessments, although he certainly did that. More critical was his theorizing about and designing of approaches to assessment that would be useful in highly complex, interactive, and adaptive domains. His work pressed the field to intentionally design assessments grounded in appropriate theories that could support learning within and from those assessments. In short, Bob's approach suggested that assessments should acknowledge that skills and understandings are situated within social contexts that can vary across individuals and settings (Mislevy, 2018). He also emphasized that stakeholders could have different purposes for using an assessment, which affects what is learned from the assessment's use (e.g., Mislevy, 2012). In this paper, we consider how Mislevy's framing and foundational contributions can be applied to the assessment of teaching.

Our approach builds from work on evidentiary arguments. Mislevy's seminal article *Evidence and inference in educational assessment* (Mislevy, 1994) laid the foundations for **evidence-centered design (ECD)**, in which he drew on philosophy, legal reasoning, and computer science to make the case for assessment being an instance of the more general problem of using evidentiary argument to make inferences in the presence of uncertainty (Mislevy, 2012, 2018). Approaching validation of teaching assessments from an evidentiary argument perspective embedded in Kane's broader validity argument (2006) approach has profound implications for what is considered relevant evidence, alternative hypotheses, and the inferences that can be supported. Any assessment of teaching must be "validated for a purpose, drawing on a specific body of evidence that considers *context-relevant* alternative explanations" (Bell & Mislevy, 2021, p. 39).

Mislevy addressed context-relevance by arguing for consideration of broader social and philosophical issues. He turned to Henry Wigmore, a scholar of legal jurisprudence, to make the point:

Wigmore concluded that there are indeed general principles to guide and analyze evidentiary reasoning, but they alone are insufficient for the full range of issues of evidence and inference that arise in jurisprudence. To begin with, questions of what constitutes evidence cannot even be framed without conceptions of the nature of people and the nature of justice (Mislevy 1994, p. 480)

These broader social and philosophical issues lie at the heart of what has been conceptualized as **culturally responsive assessment (CRA)** (Hood, 1998; Mislevy et al., 2025; Qualls, 1998). Here, we propose an approach for the design and validation of culturally responsive teaching assessments by addressing Mislevy's three central questions that provided

the foundation for all his work in assessment design: what are we trying to assess; for what purpose(s) and; how can we build an assessment that addresses these questions?

### **What is Teaching?**

Measurement broadly, and ECD specifically, require a definition of the construct being measured. This is alternatively referred to as a framework or “theory” of teaching; and it should reflect a community’s view of the construct. There is not one theory of teaching (Praetorius & Charalambous, 2023), in part because there is not one community. Rather, there are many communities of practice, often with their own theory of teaching, including those from measurement (e.g., White et al., 2024), economics (e.g., Goldhaber & Anthony, 2007) and teacher education (e.g., Shulman, 1986). Nor is there a unitary view of culturally responsive teaching. However, to develop and use culturally responsive teaching assessments we must ultimately specify what we mean by culturally responsive teaching, even if at a very high level of granularity.

**The Three Domains of Teaching.** Referred to as the instructional triangle (Cohen et al., 2003), teaching is a complex, interactional, socially situated practice that occurs between and among teachers, students, and subject matter, within varied environments. These interactions occur in three domains of teaching that have evolved from diverse literatures and disciplines (e.g., child development, educational psychology, education) in the Americas and western Europe (see also Li et al., 2020; Praetorius et al., 2018). These three domains emerge from theory-building and associated empirical studies carried out largely by educational psychologists in the mathematics education research community in Germany (see also Klieme et al., 2001) and developmental psychologists and education researchers in PreK–12 general education research in the United States (Hamre & Pianta, 2007). These “three basic dimensions” of teaching interactions have alternatively been described as cognitive activation, supportive climate (or student support), and classroom management (Klieme et al., 2001; Praetorius et al., 2018) and classroom organization, instructional support, and emotional support (Hafen et al., 2015). Following a recent global literature review of measures of teaching quality, we refer to these three domains as *classroom organization*, *supporting students socially and emotionally*, and *instructional support and activation* (Bell et al., 2025).

**Culturally Responsive Teaching.** In defining teaching for CRAs, we begin with the assumption that assessment instruments are designed to support culturally responsive teaching (e.g., Gay, 2018; Jensen et al., 2018; Lee, 1998). Decades ago, scholars began to theorize teaching, learning, and school-based experiences that would better serve students traditionally minoritized and marginalized in U.S. public schools. In the teaching field, this has evolved into rich, varied, and intersecting literatures on culturally responsive teaching (Gay, 2000), culturally responsive pedagogy (Ladson-Billings, 1995), and culturally sustaining pedagogy (Paris, 2012). Given our focus on CRA of teaching, we look across these perspectives to identify features common to all three that have direct bearing on assessments of teaching. So as not to coin a new term to the literature and follow other assessment work in this area (Franco, 2024; Ladson-Billings, 2021), we refer to them collectively as **culturally responsive teaching (CRT)**. CRT is teaching that can be conceptualized as having three domains (classroom organization, supporting

students socially and emotionally, and instructional support and activation) and has the following characteristics:

- **Asset-based Approaches.** CRT begins by recognizing, attending to, and building upon the knowledge, language, experiences and culture that each student brings to a learning environment. CRT explicitly rejects deficit models of students.
- **High Expectations for All Students.** CRT creates a learning environment in which all students can bring their full cultural repertoire, all their funds of knowledge, and engage with practices that support deep thinking, understanding, and skill demonstration.
- **Cultural Competence.** CRT prioritizes a) elevating and integrating the cultures of students into instructional practice, and b) cultural understanding and respect for students' own cultures as well as the cultures of others.
- **Critical Consciousness.** CRT supports the development of students' ability to question and analyze societal and social issues from a critical perspective, including their learning and schooling experiences.

The interactions that support CRT in the instructional triangle are, in Mislevy and Yin's (2009) terms, a complex adaptive system. Teachers do not use their knowledge and skills without accounting for the students in the room and details of the subject matter they are responsible for teaching. Further, the instructional triangle is situated in schools, in families within communities, and in nations—each of which have cultural features that impact how CRT is defined and assessed. Any approach to assessing teaching must, therefore, take account of teaching as a situated complex adaptive system that endeavors to support CRT.

### **What is the Difference Between Culturally Responsive and More Traditional Assessments of Teaching?**

Culturally responsive assessment (CRA) has been proposed as an alternative to traditional testing practices for many reasons. The first involves the lived experiences of minoritized students. As a starting point, a focus on achievement gaps and deficits, without attending to the vast disparities in resources, opportunities, and expectations that shape learning environments, has hurt marginalized students (Ladson-Billings, 2006; National Academies of Sciences, Engineering, and Medicine [NASEM], 2025). Current assessment practices focus on deficits and gaps and do not advance equity frameworks that embrace heterogeneity among students, promote justice, or envision sustainable futures (see NASEM [2025] for five frameworks of equity). Critical theorists point out that important cultural constructs have not been included in traditional assessments, e.g., sharing or helping others (Alcalá et al., 2018) and critical resistance (Ladson-Billings, 1995). And racial justice scholars have argued that assessment systems, particularly those with high stakes attached, have exacerbated racial inequities in education by exposing communities of color disproportionately to harsher consequences (e.g., Darling-Hammond, 2007; Lipman, 2002).

A second challenge to traditional assessment has to do with the relative priority that traditional assessment and CRA place on the two sources of justification for the validity of

assessments described by Messick (1989). Traditional measurement practices have given the greatest attention to the *evidential basis of testing*, focusing on construct validity and the relevance and appropriateness of test use for particular purposes. CRA puts at the forefront what Messick called *the consequential basis of testing*, focusing on the value and social implications of assessment (see also Gitomer & Iwatani, 2022).

For CRA, the consequential basis for assessment validity is that assessments support equity and social justice. Boykin (2014) noted that for educational assessments to support equity and social justice, test results need to “lead to greater (or at least equal) access to educational, professional, or status opportunities, or to at least equal performance outcomes for socially marginalized groups,” and this has to be accomplished “within the realities of a multiethnic, multiracial, and multicultural society, where status and opportunities are arrayed along ethnic, racial, and cultural lines” (p. 501; see also Lyons et al., 2025).

CRA’s critiques of traditional assessments have serious implications for culturally responsive assessment of teaching—specifically, the degree to which assessments provide information about both the individuals and systems that produce and support CRT. CRAs of teaching must focus on *teaching* rather than *teachers*. Assessments must do this because teachers—just like students—are individuals embedded in social systems. Just as a students’ social systems include their classroom, school, family, and community, a teachers’ social system includes their classroom, school, district, community, state, and nation. When assessing classroom interactions in the instructional triangle, we must pay careful attention to the entire system, not just the teacher. Teachers are employees embedded in systems and do not act independently. For example, teachers often are required to use specific curricula, and they are always enmeshed in school norms and values regarding critical aspects of classroom interactions such as discourse, achievement, cooperation, and participation. Assessments of teaching require clear thinking regarding the systems in which teachers practice. This can be accomplished through CRA design and a participatory process of validation.

Given the articulation of CRT in the previous section, CRAs of teaching should measure teaching (not teachers) and measure the core features of CRT described above.

### **What is the Purpose of Culturally Responsive Assessment of Teaching?**

If culturally responsive assessments of teaching provide information about the individuals and systems that produce and support CRT, what purposes can those assessments serve? Given the focus on equity and justice, we see two purposes of these assessments as paramount: (1) Assessing teaching to support teacher learning and (2) assessing teaching to support the flourishing of all teachers.

CRAs of teaching should have the overall purpose of supporting teacher learning. Such assessments are not designed to sort or label teachers, nor are they used as an accountability check. They are designed so that educators and education systems can iteratively and systematically improve. They are designed using the same principles used with student assessment: asset-based approaches; high expectations; and cultural competence and critical consciousness.

Because the focus is on assessing teaching, a major task is to discern patterns of teaching across classrooms. Addressing this task will identify community-wide needs that can be addressed and provide insight into situational conditions that can be altered (e.g., improved curriculum or professional development to strengthen community practices). Such an approach explicitly rejects the idea that teaching quality is a stable trait of an individual teacher independent of the context in which they teach, and that the best mechanism for improving teaching is to replace “weak” teachers with “strong” teachers, ignoring contextual conditions (e.g., Hanushek, 2002).

CRA of teaching should also support the flourishing of all teachers and communities, and particularly those that have been disenfranchised by our historical educational structures. Doing so requires explicit valuing of CRT practices and ensuring that practices that do not conform to traditional images of effective teaching are understood and not summarily rejected because of their unfamiliarity.

### **How Can We Design and Validate Culturally Responsive Assessments of Teaching?**

To achieve the goal of developing and validating culturally responsive assessments of teaching, we begin with the foundational assertions that teaching is a **situated complex adaptive system (SCAS)**, and assessments of it require a CRA approach that leverages ECD.

Existing assessments can be considered through CRA frameworks (see also Lyons et al., 2025) and validation guidance (Randall et al., 2024) that make visible potential strengths and weakness of assessment use cases, ultimately leading to revisions of those use cases and assessments. Given space considerations, we focus on the development and validation of new teaching assessments. We build on Hood (1998), Boykin (2014), Gordon (2020), Qualls (1998), and O’Dwyer et al. (2023) and use Mislevy and colleagues’ (2025) approach to designing CRA tasks and validating through an evidentiary argument, as per Kane’s (2006) view of validation.

### ***Culturally Responsive Assessments Applied to Teaching and Teachers***

CRA of teaching should themselves be culturally responsive to teachers and their associated communities. Geneva Gay draws on anthropology’s semiotic conception of culture as the “webs of significance” (Geertz, 1973) that human beings create and use to make meaning of their interactions. Gay (2010) defines *culture* as “...a dynamic system of social values, cognitive codes, behavioral standards, worldviews, and beliefs used to give order and meaning to our own lives as well as the lives of others” (Delgado-Gaitan & Trueba, 1991, p. 9). CRA of teaching should attend to the dynamic system (i.e., SCAS) that gives order and meaning to teachers and their communities.

Assessments should therefore take an asset-based approach to teachers, collecting evidence of how teachers understand their practice. Integral to that assessment, all teachers should be held to high professional standards of teaching that are collectively defined by teachers. Teachers’ cultural backgrounds vary (e.g., race, language, religion), and assessments should value the cultural competence teachers have as both professionals and as human beings. Finally, interpretations of evidence and inferences from teaching assessments should include a critical lens and serve to empower teachers to make change in their communities and profession.

***Design Principles for Culturally Responsive Assessments of Teaching***

To develop CRAs of teaching, we build on approaches taken in the student domain (see also O'Dwyer et al., 2023; Evans & Taylor, 2025). Developers should view the process as participatory, iterative, context-aligned, and context-informed.

**Participatory.** CRAs assume that social values, cognitive codes, behavioral standards, worldviews and beliefs will be specified by members of the community for which the assessment will be used. These community members will certainly include educators but may include others in the community who have a stake in teaching and learning (e.g., parents, students, administrators, and community elders). External assessment experts' role in design efforts is to offer technical expertise to support the community, not define assessment outcomes.

**Iterative.** ECD suggests the community and assessment experts must collaboratively clarify the claim(s) intended from the assessments. As community members discuss, advise, provide feedback and read and learn together (O'Dwyer et al., 2023), those claims and the specifics of the evidence necessary to support the claim(s) may change. Assessments should be used with other assessment information (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 2014) and this information may shape the design iterations.

**Context-aligned.** Traditional teaching assessment approaches are intended to produce a profile of teaching that is general across contexts, e.g., Danielson's Framework for Teaching (Danielson, 2013). While this has its place for certain purposes, to improve teaching and flourishing in local contexts, assessments will likely be aligned to a circumscribed set of goals that the community defines. In this case, evidence of competence will be updated as the community's perceived needs change.

We apply the Mislavy et al. (2025) considerations for developing situated assessments to the teaching context and contend that they have value in developing culturally responsive assessments of teaching.

- A. **Construct Element:** What are the core skills of teaching being assessed?
- B. **Sociocultural Grounding:** What is the cultural context in which the assessment task(s) is situated?
- C. **Evidentiary/Enacting Representation:** How is evidence of target teaching constructs expressed or represented in the assessment task(s)?
- D. **Ancillary Demands:** What additional skills or knowledge are required by the task that are not the primary focus (e.g., technical terminology)?
- E. **Genre/Generalization:** What is the format or social situation of the assessment (e.g., classroom observation, evaluation of classroom artifacts, student feedback)?
- F. **Other Capabilities:** What capabilities does a teacher or teachers possess that may be leveraged in an assessment but are not the primary construct (e.g., knowledge of curriculum or child development)?

- G. **Genre Demands Aspects:** What are the forms, activities, and expectations that together organize the preceding aspects into a task situation through which teachers must interact to display their capabilities?

### ***Validation of Culturally Responsive Assessments of Teaching***

Validation of CRA of teaching should adhere to Kane's (2006) validity argument approach that specifies interpretive and validity arguments are evaluated and refined over time, across multiple studies. Using the two-step approach Kane details, we must first specify the interpretive argument. In that argument, we describe the chain of inferences and assumptions that connect the performance we observed to the conclusion we wish to draw from that performance, moving from scoring to generalization to extrapolation to implication inferences. In the second step, we gather empirical evidence and additional information to construct the validity argument. This step determines the degree to which there is support for the chain of inferences theorized in the interpretive argument (for examples, see Bell et al., 2012, and Hill et al., 2011).

Traditional assessment of teaching has generally failed to employ Kane's approach by inadequately and incompletely describing the interpretive argument and inadequately and incompletely gathering evidence for each part of the interpretive argument. Particularly relevant to CRA is that stakeholder participation in the validation effort seldom extends beyond highly constrained test development processes. So, for example, validation of any CR teacher assessment would involve teachers in the community contributing to interpreting assessment evidence to make defensible interpretations and also to provide guidance on the implications of those interpretations.

Validation is a time-consuming and careful process. Frequently, policymakers, researchers, and other stakeholders are interested in whether an assessment has "enough" evidence that it "is valid," so we can begin to use it. While this impulse is understandable, it has led to inadequate validity arguments and continued harms for traditionally disadvantaged teachers and communities.

Traditional assessment has also failed to systematically engage stakeholders at every step in the process. Validation cannot stop at asking stakeholders to share their conception of the construct of teaching but not to consider the validity and meaning of assessment evidence and the resulting scores. Stakeholders can and should weigh in on each step in developing evidence for and evaluating the validity argument. Further, they should be engaged in generating alternative hypotheses and evaluating the strength of validity evidence for the interpretive argument's inferences.

A CRA approach to validation should not adopt a measurement metaphor (Mislevy, 2012). Given teaching's conceptualization as an SCAS, teaching assessments are more productively validated in Kane's comprehensive framework that uses Mislevy's evidentiary argument approach (Mislevy, 2012) for specifying and gathering empirical evidence for the validity argument step of validation. Mislevy (2019) wrote, "Rather than seeing assessment primarily as measurement, we are increasingly seeing it as an evidentiary argument, situated in social contexts, shaped by purposes, and centered on students' developing capabilities for valued activities." (p.164).

As **Figure 1** specifies, assessments that have the purpose of improving teaching and teachers' thriving require attention to each step of the evidentiary argument leading from the assessment situation to claims about the teacher. We elaborate on several parts of the argument because the history of teacher assessment has not adequately engaged a validity argument approach with stakeholders (see also National Academies, 2025; Gitomer & Marshall, 2025; Gitomer et al., 2021). Validation should pay careful attention to the a) assessment situation, b) warrant for task design, c) warrant for scoring, d) other information about the teacher in the assessment situation, and e) alternative explanations for any claims made about the teacher.

We specified the task design issues that should be addressed when creating the assessment situation and warrants for task design (i.e., a and b).

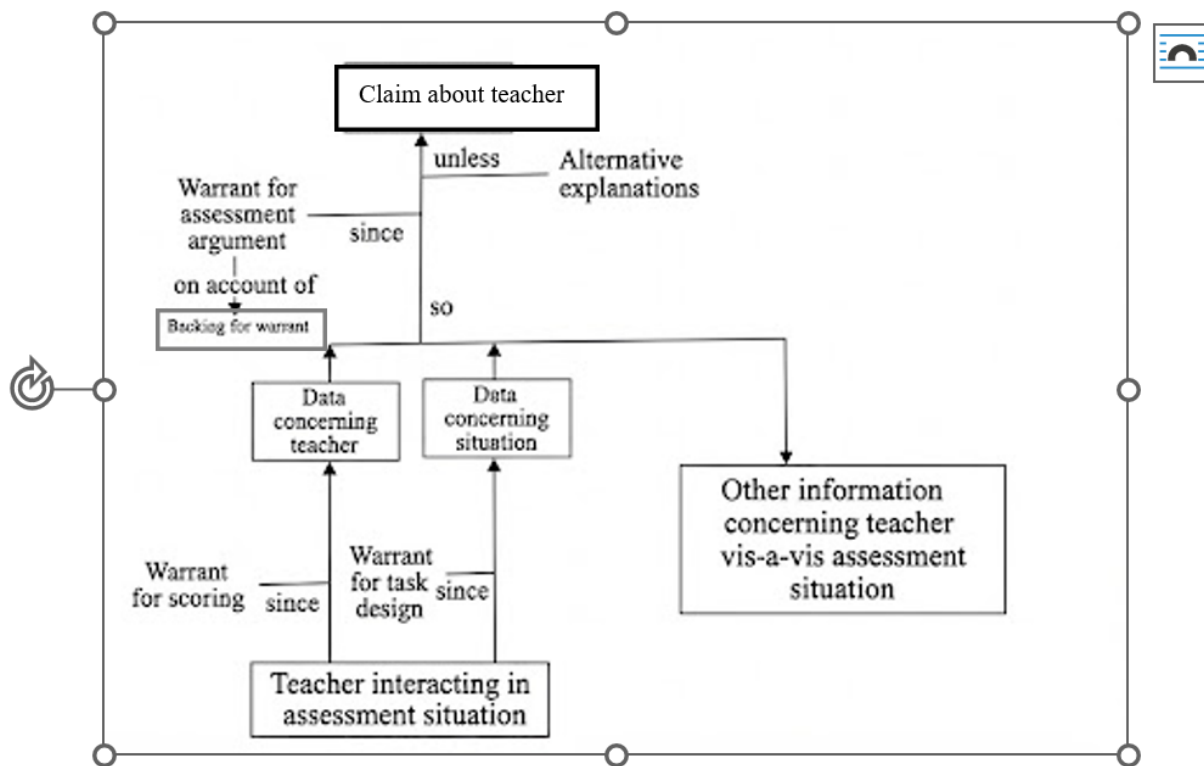


Figure 1. The structure of a teaching assessment

**Warrant for Scoring.** Given a particular assessment design, the evidence about teaching must be interpreted. Such interpretation might include traditional application of rubrics and associated scoring processes and should include solicitation of information from stakeholder that results in discussion and deliberation. In **Figure 1**, stakeholders could be engaged in developing alternative hypotheses for particular scoring patterns. They could also evaluate the warrant(s) that support the box called “data concerning teacher.” If, for example, the teacher has never had the opportunity to learn to perform the skill being assessed, this alternative hypothesis about the meaning of the teachers’ scores would undermine the warrant that should support the data for the teacher and the assessment’s claim about the teacher.

**Other Information.** Stakeholders have access to information that, if solicited and accounted for, can make for more equitable assessment. In the evidentiary argument this can be accounted for the in box specified “other information concerning teacher vis a vis assessment situation.” For example, if the assessment is misaligned to the teacher’s mandated curriculum—which she is required to know and implement every day—inferences that rely on the teacher using her curricular knowledge on the assessment may need to be reconsidered.

**Alternative Explanations.** Every inference—scoring, generalizability, extrapolation, and implication—requires the gathering and evaluation of validity evidence. These actions should engage stakeholders in actively considering alternative hypotheses for the patterns seen in that evidence. This might include the identification of and consideration of evidence about the status of existing teaching practices across the community (e.g., a school or group of schools).

## Conclusion

Both culturally responsive assessment and Mislevy’s approach to the assessment of situated complex adaptive systems raise fundamental challenges to traditional assessment practices—practices that have prioritized efficiencies to serve specific institutional purposes. Together, they recognize and emphasize the importance of developing assessments and assessment practices that attend to the complexity of human learning and development that occurs within natural, situated contexts.

For teaching, this means developing assessments that recognize that what happens between teachers and students must be understood in terms of social, cultural, and situational factors that can facilitate or inhibit teacher and student development. Culturally responsive assessments of teaching are designed to support the goals and practices of culturally responsive teaching. Effective assessment practices will support teacher learning and the flourishing of teachers within the educational system.

We see CRA and Mislevy’s work as having a symbiotic potential. CRA fleshes out the idea of situated assessments in complex assessments by explicitly addressing race, culture, language, social (in)equality and social (in)justice as issues that must be given attention. Mislevy offers approaches and tools that can enhance the potential of CRA efforts to be successful. If these two broad efforts can be brought together, the assessment of teaching will be greatly advanced.

## References

- Alcalá, L., Rogoff, B., & López Fraire, A. (2018). Sophisticated collaboration is common among Mexican-heritage US children. *Proceedings of the National Academy of Sciences*, 115(45), 11377. <https://doi.org/10.1073/pnas.1805707115>
- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <http://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition>
- Bell, C. A., Castro, M., Jörg, V., Meng, R. Xie, X., Kunter, M., & Oloff, J. (August 2025). *The validity and utility of established and emerging measures of teaching quality: A review*. [Paper presentation]. European Association for Research on Learning and Instruction (EARLI) Biannual Meeting, Graz, Austria.
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, 17(2–3), 62–87. <https://doi.org/10.1080/10627197.2012.715014>
- Bell, C. A. & Mislevy, R. (2021). Practice, feedback, argument, measurement: A frame for understanding diverse perspectives on teaching assessments. In Klette, K. & M. Tengberg (Eds.) *Analysing teaching quality: Perspectives, principles and pitfalls*. Scandinavian University Press. [https://issuu.com/universitetsforlaget/docs/sider\\_fra\\_1297-ways-of-analyzing\\_pub020222](https://issuu.com/universitetsforlaget/docs/sider_fra_1297-ways-of-analyzing_pub020222)
- Boykin, A. W. (2014). Human diversity, assessment in education and the achievement of excellence and equity. *The Journal of Negro Education*, 83(4), 499–521.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25(2), 119–142.
- Danielson, C. (2013). *The Framework for Teaching evaluation instrument, 2013 instructionally focused edition*. The Danielson Group.
- Darling-Hammond, L. (2007). Third annual Brown Lecture in Education Research—The Flat earth and education: How America’s commitment to equity will determine our future. *Educational Researcher*, 36(6), 318–334.
- Delgado-Gaitan, C., & Trueba, H. (2023). *Crossing cultural borders: Education for immigrant families in America*. Routledge.
- Evans, C. M., & Taylor, C. S. (Eds.). (2025). *Culturally responsive assessment in classrooms and large-scale contexts: Theory, research, and practice*. Taylor & Francis Routledge [Imprint].
- Franco, M. P., Bottiani, J. H., & Bradshaw, C. P. (2024). Assessing teachers’ culturally responsive classroom practice in PK–12 schools: A systematic review of teacher-,

- student-, and observer-report measures. *Review of Educational Research*, 94(5), 743–798. <https://doi.org/10.3102/00346543231208720>
- Gay, G. (2001). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.
- Geertz, C. (1973). *The interpretation of cultures: Selected essays*. Basic Books.
- Gitomer, D. H., & Iwatani, E. (2022). Two communities' views on test fairness. *Educational Assessment*, 27(2), 197–203. <https://doi.org/10.1080/10627197.2022.2087624>
- Gitomer, D. H., & Marshall, B. L. (2025). The bold and unfulfilled promises of teacher evaluation as policy. In L. Cohen-Vogel, P. Youngs, & J. Scott (Eds.). *Handbook of education policy research. 2nd edition* (pp. 1365–1390). American Educational Research Association;
- Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2021). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, 58(1), 3–31. <https://doi.org/10.3102/0002831219890608>
- Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching. *Review of Economics and Statistics*, 89(1), 134–150. <https://doi.org/10.1162/rest.89.1.134>
- Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, 39(3), 72–78. <https://doi.org/10.1111/emip.12370>
- Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the classroom assessment scoring system—secondary. *The Journal of Early Adolescence*, 35(5–6), 651–680.
- Hammond, Z. (2014). *Culturally responsive teaching and the brain: Promoting authentic engagement and rigor among culturally and linguistically diverse students*. Corwin Press.
- Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. Pianta, M. Cox, & K. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–84). Baltimore: Brookes.
- Hanushek, E. A. (2002). Teacher quality. In L. T. Izumi & W. M. Evers (Eds.), *Teacher quality* (pp. 1–13). Hoover Institution Press.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794–831. doi: 10.3102/0002831210387916
- Jensen, B., Grajeda, S., & Haertel, E. (2018). Measuring cultural dimensions of classroom interactions. *Educational Assessment*, 23(4), 250–276. <https://doi.org/10.1080/10627197.2018.1515010>
- Kane, M. T. (2006). Validation. *Educational Measurement*, 4(2), 17–64.

- Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematics teaching in lower secondary education: “Task culture” and lesson design. In E. Klieme, & J. Baumert (Eds.), *TIMSS - Impulses for school and teaching* (pp. 43–57). Federal Ministry of Education and Research.
- Kraft, M. A., Papay, J. P., James, J. K., & Monti-Nussbaum, M. (2026). *Is teacher effectiveness fully portable? Evidence from the random assignment of transfer incentives* (NBER Working Paper 34845). <https://doi.org/10.3386/w34845>
- Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal*, 34(1), 39–81.
- Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3), 465–491.
- Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, 35(7), 3–12. <https://doi.org/10.3102/0013189X035007003>
- Ladson-Billings, G. (2021). *Culturally relevant pedagogy: Asking a different question*. Teachers College Press.
- Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education*, 67(3), 268–279. JSTOR. <https://doi.org/10.2307/2668195>
- Li, H., Liu, J., & Hunter, C. V. (2020). A meta-analysis of the factor structure of the Classroom Assessment Scoring System (CLASS). *The Journal of Experimental Education*, 88(2), 265–287. <https://doi.org/10.1080/00220973.2018.1551184>
- Lipman, P. (2002). Making the global city, making inequality: The political economy and cultural politics of Chicago school policy. *American Educational Research Journal*, 39(2), 379–419. <https://doi.org/10.3102/00028312039002379>
- Lyons, S., Oliveri, M. E., & Poe, M. (2025). A framework for enacting equity aims in assessment use: A justice-oriented approach. In *Culturally responsive assessment in classrooms and large-scale contexts* (pp. 88–105). Routledge.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). American Council on Education.
- Mihaly, K., & Mccaffrey, D. F. (2015). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 7–49).
- Mislevy, R. J. (2012). *Four metaphors we need to understand assessment*. The Gordon Commission on the Future of Assessment in Education. [http://www.gordoncommission.org/rsc/pdf/mislevy\\_four\\_metaphors\\_understand\\_assessment.pdf](http://www.gordoncommission.org/rsc/pdf/mislevy_four_metaphors_understand_assessment.pdf)

- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mislevy, R. J. (2019). Advances in Measurement and Cognition. *The ANNALS of the American Academy of Political and Social Science*, 683(1), 164–182.
- Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning*, 59, 249–267. doi:10.1111/j.1467-9922.2009.00543.x
- Mislevy, R. J., Oliveri, M. E., Slomp, D., Wolf, A. C. E., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. (pp. 199–241) In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. Taylor & Francis.
- National Academies of Sciences, Engineering, and Medicine. (2025). *Equity in K-12 STEM education: Framing decisions for the future*. The National Academies Press. <https://doi.org/10.17226/26859>
- O'Dwyer, E. P., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Applied Measurement in Education*, 36(3), 286–303. Academic Search Premier (164367266). <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>
- Praetorius, A. K., & Charalambous, C. Y. (2023). Where are we on theorizing teaching? A literature overview. In A.-K. Praetorius & C. Y. Charalambous (Eds.), *Theorizing teaching: Current status and open issues* (pp. 1–22). Springer International Publishing. [https://doi.org/10.1007/978-3-031-25613-4\\_1](https://doi.org/10.1007/978-3-031-25613-4_1)
- Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM–Mathematics Education*, 50(3), 407–426.
- Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2024). Justice-oriented, antiracist validation: Continuing to disrupt white supremacy in assessment practices. *Educational Assessment*, 29(1), 1–20. <https://doi.org/10.1080/10627197.2023.2285047>
- Qualls, A. L. (1998) Culturally responsive assessment: Development strategies and validity issues. *The Journal of Negro education*, 67(3), 296–301.
- Schwab, J. J. (1978). *Science, curriculum, and liberal education: Selected essays*. University of Chicago Press.
- White, M., Edelsbrunner, P. A., & Thurn, C. M. (2024). The conceptualisation implies the statistical model: Implications for measuring domains of teaching quality. *Assessment in Education: Principles, Policy & Practice*, 31(3–4), 254–278. <https://doi.org/10.1080/0969594X.2024.2368252>