CASE STUDY

Conceptualizing and Evaluating Instructionally Useful Assessments

Scott F. Marion and Carla M. Evans

Conceptualizing and Evaluating Instructionally Useful Assessments

Scott F. Marion and Carla M. Evans

It seems silly to ask students to take a test that is not useful for improving their learning. However, many tests are designed to serve purposes other than instruction, such as monitoring and evaluating educational programs and school quality. Even still, far too many test vendors claim their tests are instructionally useful even if they were designed to serve other primary purposes. We rarely see evidence supporting these claims. In particular, we are concerned that teachers are often blamed for not using assessment results to improve their instruction. Doing so with assessments not designed to support instructional inferences is a Sisyphean task. Therefore, we wrote *Understanding Instructionally Useful Assessment* (Evans & Marion, 2024) to clarify our perspective on what it might take for an assessment to provide instructionally useful information.

Not surprisingly, our interest in defining and clarifying instructional usefulness parallels efforts to rethink balanced assessment systems more broadly. In the recent National Academy of Education publication, *Reimagining Balanced Assessment Systems* (Marion et al., 2024), the authors focus on rebalancing assessment systems to privilege rich classroom learning environments. The first part of the updated definition makes this point:

> *Balanced assessment systems and practices, as conceived by this volume's authors, are intentionally designed to provide feedback to students and information for teachers to support ambitious and equitable instructional and learning opportunities. This type of assessment system facilitates educator engagement in high-leverage professional practices such as quality formative assessment to support ambitious and equitable teaching* (Marion et al., 2024, p. 2).

As the definition shows, assessments that support learning and instruction are the focal point of a balanced assessment system. The authors emphasized that high-quality formative assessment practices can best support instructional uses. But what does it take for an assessment to be intentionally designed and implemented to support rich learning environments by providing instructionally useful information?

In the sections that follow, we first define instructionally useful assessments, then discuss key assessment design features that facilitate instructional usefulness, and conclude with some evidence requirements and areas for future research.

Instructional Usefulness Defined

> *"Although there is often a big difference between identifying a weakness and correcting it, identification can be a major part of the battle. By itself, a test score does little to identify the nature of a problem, only that there is one"* (Linn, 1983, p.179).

We started our journey into instructional usefulness by revisiting the wise words of some of the giants in our field. In addition to Robert Linn, assessment luminaries such as Eva Baker, Joan Herman, Peter Airasian, and many others contributed articles to a special issue in the *Journal of Educational Measurement* in 1983. However, even before 1983, assessment professionals have been concerned about having assessments support meaningful instructional actions.

We consolidated the ideas of many who came before us and defined an instructionally useful assessment as one that "…provides substantive insights about student learning strengths and needs relative to specific learning targets that can positively influence the interactions among the teacher, student, and the content" (Evans & Marion, 2024, p. 19). We further explored how instructionally useful assessments can support teachers by revealing insights through the assessment processes themselves, reporting results that shed light on student learning, or simply as a function of participating in the assessment (e.g., Agarwal et al., 2008).

Our definition, particularly the notion of substantive insights, follows from the conceptualization of the "assessment triangle" as described in the seminal volume, *Knowing What Students Know* (NRC, 2001). The authors emphasized that such insights are most likely to occur when there is coherence among the learning goals and progressions, the assessments (observations), and the interpretative approaches. These substantive insights occur when the assessments support interpretations in light of the progressions by which students are expected to achieve domain competence.

An instructionally useful assessment provides substantive insights about student learning strengths and needs relative to specific learning targets that can positively influence the interactions among the teacher, student, and the content.

We are aware that some might find this conception exclusionary. That was our intent. We have been concerned about claims that state standardized tests and commercial interim assessments can provide instructional insights, leading to considerable confusion about what it might take for an assessment to support instruction. We found Elmore's (2008) conception of the "instructional core" critical for framing our argument that instructionally useful assessment plays out in the interactions among students, teachers, and rich content within engaging learning environments.

Optimizing the conditions that make assessments more or less instructionally useful is necessary but insufficient for producing instructionally useful insights. We relied on Elmore (2008), Faxon-Mills et al. (2013), and Coburn and Turner (2011) to help us conceptualize how mediating variables influence instructional utility. In the case of instructional usefulness, the teacher's interpretive processes, decision-making, and instructional repertoires are critical mediating factors in influencing the instructional usefulness of assessments.

Beyond the skills of the teachers interpreting the assessment activities, other mediating factors include selecting an assessment that matches the enacted curriculum, having score reports that support productive actions, and ensuring that teachers have time and capacity to review and act on the assessment results. As we described in our book, the challenges of mediating factors increase as the assessment moves further away from the direct interaction between teachers and students. The presence of mediating factors is not an excuse for why an assessment does not yield instructionally viable insights. We highlight the importance of mediating factors so assessment designers and education leaders can understand the additional requirements and steps necessary for assessments to serve instructional purposes.

Design and Implementation Features

We spent considerable time thinking about the features of an assessment and the ways it is implemented and scored that would increase or decrease the likelihood that the assessment results could support instructionally useful interpretations and actions. We identified the following ten features as most relevant to explain why some assessments, and the information they produce, are useful for directly informing instruction, while others are less so.

1. Cognitive complexity and associated item types
2. Coherence with the enacted curriculum
3. Breadth of content standards and resulting grain size of results
4. Type of results (quantitative/qualitative)
5. Timing of results (e.g., ongoing, weekly, yearly)
6. Administration and scoring conditions
7. Allowable student responses
8. Student choice
9. Collaboration
10. Real-world and culturally relevant connections

Lorrie Shepard (2024) suggested that we organize the 10 features into three major categories. We provide additional details for these features, as categorized by Shepard, in the paragraphs that follow.

Representation of the Learning Goals

The ways in which the learning goals are conceived and represented are the most critical for influencing the potential instructional usefulness of an assessment and the associated results. We posit that coherence with the enacted curriculum, cognitive complexity of the test items, and breadth of content standards strongly influence the potential instructional usefulness of an assessment. This concept harkens back to Wiggins and McTighe's (2011) notion of backward design, where instructional designers are asked to begin with clear, desired learning outcomes and then design evidence of learning related to these intended outcomes. If an assessment is not embedded in the curriculum being taught, the teacher is forced to undertake several additional interpretative steps to make sense of the curriculum-agnostic results in light of their curriculum.

As we explained above, deriving substantive insights into current states of student learning requires understanding students' location on a progression of learning rather than whether they have grasped it or not. Therefore, instructionally useful assessments must include a range of item types to probe different levels of cognitive complexity. Finally, instructional insights are generally tied to specific aspects of the curriculum and content standards. Therefore, assessment results should be at a small enough grain size so teachers can understand how to focus on specific knowledge and skills (Evans & Marion, 2024; Marion et al., 2024).

Just-in-Time Insights into Student Thinking

The timing and type of results can influence the instructional usefulness of assessments. At its simplest, results should be returned when it makes sense to offer "just-in-time" instructional adjustments so students can competently engage in the next curricular unit. Additionally, teachers need opportunities to "see" student thinking in order to derive substantive insights. Quantitative results can be challenging to interpret due to the obscure ways they are presented, especially for those without a deep quantitative background. We recognize the challenges associated with presenting actual student work for every response, but including at least some descriptive representations can help the interpretability of the full set of test results.

We also suggest that the degree of standardization required (or flexibility allowed) regarding the administration requirements, scoring conditions, and allowable student responses can affect instructional usefulness. These features are not as critical as the five features already discussed, but we believe these three features can influence the utility of the assessment results. Highly standardized administration and scoring requirements, as well as restricted types of allowable student responses, can limit students' ability to demonstrate their knowledge and skills. This might hinder teachers' understanding of what their students know and what they need to learn next to progress in their learning.

Sociocultural/Affective Aspects of Student Learning

We grounded our conceptualization of instructional usefulness in sociocultural learning theory and ambitious teaching (e.g., Ball & Forzani, 2009; Shepard, 2021). We are less certain about the degree to which student choice, collaboration, and real-world and culturally relevant connections influence instructional usefulness, but they are coherent with our theoretical orientation. Student choice could include flexibility in student learning goals, assessment targets, assessment methods, and flexibility in ways of demonstrating one's learning. Adults (and students) are expected to collaborate in cultural communities of practice, yet most of our assessments isolate students from these communities. Would teachers be able to derive more useful instructional insights if students were assessed in ways that more authentically allow them to show what they know and represent the world and their cultural communities? We suspect so.

Summary of Design and Implementation Features

The ten assessment design and implementation features all exist on a continuum. When most of the features, especially the first seven, are represented in ways to support substantive insights about student learning, the assessment is more likely to be instructionally useful. The last three features support instructional use because they support ambitious teaching practices and are often related to many other assessment features. For example, high-quality performance tasks can often support increased student choice, collaboration, and culturally relevant connections. Such tasks can generally yield descriptive results that provide insights into students' thinking.

It is not enough for only one of these features to be present. Many of the highest leverage features must be present because they mutually support one another. For example, assessment results that facilitate educators' insights into students' thinking are necessary, but insufficient to support instructional usefulness. If the assessment results do not provide substantive insights for teachers when they could feasibly adjust their instructional approach (e.g., timing is off) or teachers do not understand how the results relate to their enacted curriculum, the results will likely be of little instructional value.

While we can optimize features that make assessments more instructionally useful, assessments do not operate in a vacuum. For example, a high-stakes accountability use case would likely overshadow any potential instructional usefulness, even if the test was designed and implemented with many positive features (Evans & Marion, 2024; Marion et al., 2024).

Evidence of Instructional Usefulness

Claims are statements about what a product, person, or process will do under certain conditions. Claims are hypotheses that must be evaluated with evidence. Unfortunately, very little public evidence exists to support claims of instructional usefulness often made by test vendors who sell and education leaders who purchase tests (Diggs, 2019; Hill, 2020; Perie et al., 2009). What types of evidence would convince us that an assessment is instructionally useful? We describe four types of studies that may provide evidence to support or refute claims of instructional usefulness.

Instructional usefulness isn't a feature to market, it's a claim that must be designed for and proven with evidence.

These studies are all somewhat post-hoc. That is, one needs the assessment and assessment items in order to conduct these studies. However, we argue that claims and evidence associated with instructional usefulness should start in the design phase. This idea is not new. It is at the heart of Evidence-Centered Design (Mislevy & Riconscente, 2006), which provides a formal framework for defining the claims about student knowledge, the observations (tasks) that would elicit evidence for those claims, and the interpretative models to evaluate the evidence.

- **Efficacy studies based on randomized controlled trials (RCTs)** are the "gold standard" of research; however, they are complicated to conduct in education. RCTs generally involve randomly selecting a sample of teachers from the population, randomly assigning a set of teachers to the "treatment" group (e.g., they get the assessment under question), randomly assigning another set of teachers to the "control" group, and then identifying a meaningful outcome variable and evaluate the difference between the two groups on this outcome. Despite the "gold standard" label, we believe RCTs are impractical, fail to account for considerable contextual influences, and overlook key insights from other approaches.
- **Cognitive laboratories**, also known as think-aloud protocols, ask participants to verbalize their thinking as they engage in an activity. We do this with students to understand how they engage with test items in the test-development stage. We can gain similar insights from teachers as they interact with student work and assessment score reports. In these studies, teachers are prompted to interpret student work or score reports as if they are talking to another teacher or if they are lesson planning. The interviewer probes for descriptions of specific interpretations the teacher generates from student work or score reports. The researcher also asks teachers to describe their specific instructional actions based on their interpretations.
- **Classroom observations** conducted by well-trained observers can shed light on how teachers make sense of assessment information. Compared with other methods, classroom observations have a major advantage: they provide insight into how teachers interpret and act on formative assessment activities and other informal assessments during instruction. Classroom observations also allow us to gather data about the effectiveness of teachers' actions based on their interpretations of assessment results and provide an important counterpoint to the types of interpretations and actions from summative or interim assessment activities.
- **Surveys** can provide information about instructional usefulness, but they require carefully crafted questions that pose scenarios to elicit evidence of how teachers interpret student work samples and score reports. Surveys that ask teachers if they found the assessment results instructionally useful are not worth the effort because of socially desirable responses. An advantage of surveys is that researchers can collect data from a representative sample of respondents, allowing for types of generalizations that are difficult to accomplish with smaller-scale studies, such as cognitive laboratories and observations. But, again, they have to be thoughtfully designed.

Evaluating Evidence

Once the data are collected, researchers should be able to indicate whether and how the assessment supported instructionally useful interpretations and actions. For example, teachers' interpretations and actions based on curriculum-embedded formative assessment opportunities could be compared to the interpretations and actions teachers derive from assessments further from instruction and the enacted curriculum to evaluate claims about the potential instructional usefulness of a particular assessment.

However, teachers are not blank slates. They interpret new assessment information in light of what they already know about their students, including their learning strengths and needs in specific content areas. Therefore, evaluating evidence of instructional usefulness must be contextualized in terms of what teachers already know. If the additional assessment does not provide useful and usable insights, users and decision-makers must consider whether it is worth the time to administer an additional assessment if it only provides redundant insights.

Responsibility for Collecting Evidence

We argue that those making claims are responsible for providing the evidence to support them. Even less formal claims, such as those found on websites or other marketing materials, necessitate supporting evidence. We acknowledge that collecting evidence to evaluate the claims requires time and resources, but the types of studies we described above are not overly challenging to conduct.

While test vendors are primarily responsible for collecting evidence to support their claims, those making assessment decisions (e.g., district leaders) are also responsible. When shopping for assessments, district leaders and other decision-makers should ask for evidence of instructional usefulness before making a purchase. Evidence could come from the types of studies we suggested. Additionally, sample student work products or score reports resulting from an assessment's administration could allow district leaders to collaborate with their teachers to evaluate the extent to which the assessment results support instructional decisions and actions in their specific context.

Future Research

Collecting the types of evidence described above will help further our understanding of the features and characteristics that contribute to an assessment's instructional utility. However, additional research can help systematically examine various aspects of instructional usefulness.

We posited ten features that influence the likelihood of an assessment yielding instructionally useful information. We based our selection on what we could glean from the literature and our years of working closely with teachers to support their assessment literacy. However, systematic research is necessary to provide insight into which features are more or less critical. Such research can shed light on the criticality of specific features (e.g., being able to examine student work) and the degree to which other features can compensate for shortcomings in other features.

Furthermore, rigorous research can help us understand how the various features interact with one another, as well as teacher expertise, school structure and culture, and other relevant factors. For example, we suspect that teachers with lower levels of assessment experience would be more likely to use assessment results if they are in schools with positive assessment cultures where they are provided time and support to interpret and act on results, compared to being in schools with weaker assessment cultures.

Finally, we need considerably more research into how score reports from standardized tests facilitate or hinder score interpretation and use. Most standardized assessments that teachers and students experience do not include descriptive or qualitative results in a way that easily supports inferences about student thinking. The quantitative results are generally presented in a formal score report. There has been extensive research into score reports (e.g., Hambleton & Zelinski, 2013), but not necessarily research into how score reports facilitate or hinder specific types of instructional inferences and actions beyond (re)grouping students.

Closing

Assessing students to gather information necessary for improving teaching and learning makes so much sense, it would be fair to question why we needed to write this chapter, let alone an entire book. Unfortunately, teachers and leaders are inundated with data, and there are many reasons why it is hard to translate the snowstorm of data into useful information. We need more research, but we hope we have helped conceptualize instructional usefulness in ways that motivate assessment designers and users to more seriously attend to what it takes to best support learning and instruction.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger III, H. L., & McDermott (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22, 7*, 861–876.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60, 5*, 497–511. https://doi.org/10.1177/0022487109348479

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement*, 9: 173–206.

Diggs, C. (2019, September 19). *Interim assessment? Didn't you mean formative assessment?* Center for Assessment CenterLine Blog. https://www.nciea.org/ blog/interim-assessment-didnt-you-mean-formative-assessment/

Elmore, R. F. (2008). *Improving the Instructional Core.* https://achievethecore.org/content/upload/Improving%20The%20Instructional%20Core_Elmore%20Article.pdf

Evans, C. M., & Marion, S. F. (2024). *Understanding Instructionally Useful Assessment*. Routledge.

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New Assessments, Better Instruction? Designing Assessment Systems to Promote Instructional Improvement.* Santa Monica, CA: RAND Corporation, 2013. https://www.rand.org/pubs/research_reports/RR354.html

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.). *APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association. https://doi.org/10.1037/14049-023

Hill, H. (2020, February 7). Does studying student test data really raise test scores? *Education Week*. https://www.edweek.org/leadership/opinion-does-studying-student-data-really-raise-test-scores/2020/02

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20, 2*, 179–189.

Marion, S. F., Pellegrino, J. W., & Berman, A.I. (2024a). Reimagining balanced assessment systems: An introduction. In Marion, S. F., Pellegrino, J. W., & Berman, A.I. (Eds.), *Reimagining Balanced Assessment Systems*. Washington, DC: National Academy of Education.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024b). *Reimagining Balanced Assessment Systems*. Washington, DC: National Academy of Education.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Lawrence Erlbaum Associates.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (J. Pellegrino, R. Glaser, & N. Chudowsky, Eds.). National Academy Press.

Perie, M., Marion, S. F., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28,* 3, 5–13. https://doi. org/10.1111/j.1745-3992.2009.00149.x

Shepard, L. A. (2021). Ambitious teaching and equitable assessment. *American Educator, 45,* 3.

Shepard, L. A. (2024, April 14). *Discussant comments. Is your test instructionally useful? How do you know?* [Symposium]. The Annual Meeting of the National Council of Measurement in Education. Philadelphia, PA.

Wiggins, G., & McTighe, J. (2011). *The Understanding by Design guide to creating high-quality units*. ASCD.

About the authors

Scott F. Marion, Ph.D., is a principal learning associate at the National Center for the Improvement of Educational Assessment. He is a national leader in conceptualizing and designing innovative and balanced assessment systems to support instructional and other critical uses. He has also led extensive work across the country to design and implement school accountability systems. Scott is an elected member of the National Academy of Education and is one of three measurement specialists on the National Assessment Governing Board, which oversees the National Assessment of Educational Progress. He coordinates and/or serves on 10 state or district technical advisory committees for assessment and accountability. He has served on multiple National Research Council committees, including those that provided guidance for next-generation science assessments, investigated the issues and challenges of incorporating value-added measures in educational accountability systems, and outlined best practices in state assessment systems. Scott is a co-author of the validity chapter in the 5th edition of *Educational Measurement,* a co-editor of the National Academy of Education's *Reimagining Balanced Assessment,* and a co-author of *Instructionally Useful Assessment.* He has published dozens of articles in peer-reviewed journals and edited volumes, and he regularly presents his work at the national conferences of the American Educational Research Association, National Council on Measurement in Education, and the Council of Chief State School Officers. Scott earned a Ph.D. from the University of Colorado Boulder with a concentration in measurement and evaluation.

Carla M. Evans is an associate director at the National Center for the Improvement of Educational Assessment. Carla supports states in designing and implementing assessment and accountability reforms. Her work includes leading statewide assessment system reviews, assessment literacy initiatives, and performance assessment design and implementation activities. Carla's research focuses on the impacts and implementation of assessment and accountability policies on teaching and learning. She is especially interested in policy research related to balanced assessment systems, culturally responsive assessment, performance-based assessments, AI in classroom assessment, instructionally useful assessment, and assessment literacy. Carla recently published two books. The first is with Scott Marion, *Understanding Instructionally Useful Assessment,* which details the assessment design and implementation features necessary to make it more likely that classroom teachers can use the information from an assessment to change their instruction the next day or within a short amount of time. Carla also co-edited an NCME volume with Catherine Taylor, *Culturally Responsive Assessment in Classrooms and Large-Scale Contexts: Theory, Research and Practice.* Carla has published numerous articles in peer-reviewed journals and regularly presents her research at the American Educational Research Association (AERA), National Council for Measurement in Education (NCME), and the National Conference on Student Assessment (NCSA). Carla received a Ph.D. from the University of New Hampshire with a concentration in Assessment, Evaluation, and Policy.

About the Study Group

The Study Group exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy; uncover future design needs and opportunities for educational systems; and generate recommendations to better meet the needs of students, families, and educators.

Date of Publication
April 2026

Citation
Marion, S. F., & Evans, C. M. (2025). Conceptualizing and evaluating instructionally useful assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning.* University of Massachusetts Amherst Libraries