

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/324214256>

The Opportunities and Challenges of a Systems Approach to Assessment

Article in Educational Measurement Issues and Practice · April 2018

DOI: 10.1111/emip.12193

CITATIONS

8

READS

1,039

1 author:



Scott Marion

27 PUBLICATIONS 960 CITATIONS

SEE PROFILE

The Opportunities and Challenges of a Systems Approach to Assessment

Scott F. Marion, *National Center for the Improvement of Educational Assessment*

Creating balanced assessment systems is really hard. There are few examples of well-functioning assessment systems, other than a limited number of research–practice partnerships. Marc Wilson’s recent presidential address to the National Council of Measurement in Education (Wilson, 2018) and Shepard, Penuel, and Pellegrino (2018) extend our understanding of assessment systems by building on conceptualizations from *Knowing What Students Know* (NRC, 2001). Both papers make the case that high-quality classroom assessments should be situated within balanced assessment systems located at district or state levels. I first discuss slightly the varied orientations toward classroom assessment expressed in these two papers and then focus most of my efforts on practical suggestions for moving productively forward on the development and implementation of balanced assessment systems.

Assessment to Support Learning

Both papers highlight the criticality of high-quality classroom assessment for producing actionable information that educators can use to improve student learning. As leaders in both formative and classroom assessment, these authors address a common misconception that classroom and formative assessment are one and the same, yet both note that high-quality formative assessment must be a key part of a classroom assessment system. Both papers emphasize that learning progressions are the “vertical” connection between large-scale and classroom assessments and the “horizontal” glue among standards, curriculum, and assessment. Both authors discuss the importance of the curriculum–assessment connection rather than just the standards–assessment link, but in somewhat different ways. Wilson (2018) argues that while assessment results “can be interpreted directly in terms of those curriculum constructs,” such interpretations are best facilitated through the “construction of learning progressions” (p. 11). On the other hand, Shepard et al. (2018) emphasize the importance of curriculum specificity to provide an interpretative framework for both instruction and assessment.

Wilson (2018) works from a principled assessment design framework and brings measurement expertise and methods to support improvements in classroom assessment. This orientation is seen throughout the paper and is perhaps most exemplified by his discussion of the measurement model as the “principled way to use the data from the students’ responses to the items, as represented in the outcome space to place the students and the items along the construct map”

(p. 27). Shepard et al. (2018) agree that measurement experts could contribute meaningfully to classroom assessment, but they caution that “measurement frameworks could distort high quality instruction, if the emphasis was on the quantification rather than the qualities of student thinking” (p. 1). These two statements appear to indicate fundamental differences in orientation. In spite of these differences, there is significant overlap among both papers and they offer aspirational visions of classroom assessment. So aspirational, in fact, I question how either of these visions can be implemented at scale. Yes, both authors provide exciting examples of their work in practice but thus far only as part of intense research–practice partnerships.

Systems Thinking

Both Wilson (2018) and Shepard et al. (2018) discussed the importance of connecting large-scale and classroom assessments through research-informed learning progressions. These papers focused justifiably on classroom assessment because their main use case is improving learning but also on raising awareness among professional measurement experts about how they might contribute to improving classroom assessments. Both papers recognize the role, albeit limited, of large-scale tests. Nevertheless, the Every Student Succeeds Act makes clear that state accountability tests are here to stay. Even though the current generation of statewide assessments includes items and tasks to better probe the depth of students’ thinking, educators and the public have not embraced these better tests. No doubt, a significant part of this rejection is due to the policy uses of the tests, but it is also because educators understand that the tests are far too distal from instruction, at the wrong grain size, and administered at the wrong time of year to make a difference in their daily practice. It has become increasingly clear that the multiple users of assessment information need balanced systems of assessments to both support learning and other purposes such as accountability.

Assessment systems are balanced when the various assessments in the system are coherently linked, often through a clear specification of the learning targets, comprehensively support multiple purposes and uses, and continuously document student progress over time. These properties of coherence, continuity, and comprehensiveness originally described in *Knowing What Students Know* (NRC, 2001) help create a powerful image of a high-quality system of assessments. Building from NRC (2001), we have found that coherence, utility, and efficiency are a bit more practitioner-oriented when working with district and state leaders (Chattergoon, 2016; Chattergoon & Marion, 2016).

Scott Marion, Executive Director, National Center for the Improvement of Educational Assessment, 31 Mount Vernon St., Dover, NH 03820; smarion@nciea.org.

Coherence

Drawing from *Knowing What Students Know* (NRC, 2001), a coherent assessment system must be compatible with the ways in which student learning is expected to progress in domain. Both Shepard et al. (2018) and Wilson (2018) discuss the ways in which learning progressions or learning trajectories should serve as the organizing framework connecting the various assessments and learning activities in a vertically coherent system. Developing learning progressions at a fine enough grain size to support assessment design and use has been challenging, and we have limited examples of where this has been done well. The call for *curricular specificity* is an important contribution from Shepard et al. (2018), because, as they note, curriculum and associated assessments “are the means by which theories of learning come to be enacted in classrooms and potentially could be made coherent across levels of the system” (p. 3).

Utility

Assessment systems are employed to serve multiple and often diverse purposes; therefore, a key criterion for assessment system quality should be the degree to which the system provides the information necessary to support the intended aims. Utility cannot be evaluated in the abstract, but it must follow from a well articulated theory of action that specifies the various intended outcomes for the system and the processes and mechanisms by which these outcomes will be realized (e.g., Hall, 2014). Both papers offer concrete examples of high-utility classroom assessment systems.

Efficiency

Efficiency means getting the most out of assessment resources and eliminating redundant, unused, and untimely assessments. Efficiency evaluations therefore should identify and reduce assessments that are not serving the stated purposes or are redundant with other, more useful assessments. Unfortunately, many district personnel assume that a set of assessments is a system if it contains at least summative, interim, and formative components. There is no magic about having interim assessments as a required component of balanced assessment system. In fact, many commercial interim assessments may distract educators from the rich assessment opportunities articulated by Wilson (2018) and Shepard et al. (2018).

Moving into Practice

Defining criteria has been critical for conceptualizing and offering a vision for assessment systems that can advance student learning in ways that both Wilson (2018) and Shepard et al. (2018) describe. Further, while both papers offer powerful examples for effective classroom assessment systems, we are still left wanting for models of coherent, efficient, and useful assessment systems that go beyond the classroom. Like Shepard et al. (2018), I have argued (Marion, 2015) that districts are the appropriate organizational level for instantiating balanced systems of assessment. In the section that follows, I offer several considerations for helping districts design and enact such systems. States, in general, are the wrong entity for the development of balanced assessment systems, but states can play a role in supporting high-quality

assessment systems. Following the discussion of the district role, I offer some promising examples of ways that states are helping support more educationally effective systems of assessments.

Districts as Loci of Control

One of the major issues with developing a coherent assessment system is figuring out who is in control. Most states are “local control,” some more so than others. This means that the state gets to control the state end-of-year assessment, but little else. In many states, having the state implement additional assessments beyond the end of year, such as PARCC’s attempt at “through-course” assessment, is seen as an assault on local control of curriculum. There is considerable rhetoric regarding the need for state-led comprehensive systems, and the state has a need to control its accountability test, but districts want to control certain district-wide assessments, and schools own even finer grained assessments. Importantly, teachers are responsible for most classroom assessments in order to best meet the instructional needs of students. Implementing balanced assessment systems cannot be a solely state-driven enterprise and, in addition to critical technical and conceptual details discussed by Wilson (2018) and Shepard et al. (2018), the tenuous political and ownership boundaries cannot be ignored.

Implementing learning progressions through the creation of high-quality curricular materials should be a district initiative (Shepard et al., 2018). Therefore, districts need to be the loci of control for balanced assessment systems. Depending on the district/school relationships, district offices tend to have at least a say in almost all assessment decisions other than formative and perhaps other classroom assessments. There is no question that an onerous state assessment (and accountability) system might negatively influence a district’s capacity to implement a high-quality assessment system, yet a coherent district system could serve as a buffer to a weak state system.

Unfortunately, most district assessments are a poorly articulated mix of legacy assessments and “multiple measures” cobbled together into an overwhelming and often incoherent picture of student learning. Much of the assessment proliferation at the district level is a result of historical programs that maintain assessments that might have been useful in the past but never seem to get retired. However, one of the major causes of incoherent district assessment systems is the massive increase in interim assessments during the No Child Left Behind era and continuing today (Perie, Marion, & Gong, 2009). Districts (and states) are flooded with offers from assessment vendors that promise to improve student learning. Not all of these programs are low-quality and ineffective, but many are (Konstantopoulos, Miller, van der Ploeg, & Li, 2016). Unfortunately, because of the low cognitive demand called for by many interim assessments (e.g., Li, Marion, Perie, & Gong, 2010), such assessment results likely distract educators from a deeper learning agenda.

However, districts do not implement these programs in an effort to waste money. They do so because they think that such assessments are a critical component of an assessment system and they are struggling to find any ray of hope to improve the performance in situations with scarce resources. There is also a belief that test results from an external entity are “official.” Interim assessments may fit within a district’s theory of action, but needing an “official” score should not

count as a legitimate reason even if some leaders question teacher-generated information. Further, the approaches put forth by Wilson (2018) and Shepard et al. (2018) should help bridge this perceived credibility gap between teacher assessments and commercial products.

States Have a Role: Tight and Loose Coupling

In an effort to extend the discussion about state-level assessments in Wilson (2018) and Shepard et al. (2018), I highlight the ways that states can play a meaningful role in supporting systems of assessment. The criteria outlined in *Knowing What Students Know* (NRC, 2001) and further developed by Chattergoon and others (Chattergoon, 2016; Chattergoon & Marion, 2016) are based on visions of *tightly coupled* systems with information flowing among the various components to maximize efficiency and utility. This is a high bar and, based on the lack of real-world examples, likely beyond the current reach of most educational systems. Recent work on designing assessments to evaluate student learning of the Next Generation Science Standards (NGSS) (Marion & Penuel, 2017; NRC, 2014) leads me to consider *loosely coupled* systems. Such systems have multiple levels of assessments tied to the same learning targets and vision of learning science to at least partially address the coherence criterion. However, because the information would not be shared across levels of the system, such loosely coupled systems would not be as efficient as ones where information from one level (e.g., classroom) could be used to support purposes at another level (e.g., accountability).

Several states are beginning to implement such loosely coupled systems of assessment. Delaware (McCrae, 2017) has historically relied on a state–district partnership to develop science curriculum materials to support instruction. The state is capitalizing on this “common” curriculum to implement common end-of-unit assessments three times per year. These assessments, relying on rich performance tasks as well as other item types, will be scored locally to serve important classroom assessment functions. They will not be used to support accountability. Rather, in select grades, the state will administer *integrative transfer tasks* to meet federal reporting and accountability purposes. The state knows that it is potentially leaving some technical quality “on the table,” especially in terms of alignment and generalizability by not including the results from the unit-based assessments as part of accountability, but is unwilling to risk the potential distortion of purposes.

Kentucky does not have the type of common curriculum found in Delaware but is employing a systems approach to science assessment with multiple components designed to serve different users (Kentucky Department of Education, 2017a). Following the recommendations put forth in *Developing Assessments for the Next Generation Science Standards* (NRC, 2014), the Kentucky Department of Education (KDE) is distributing high-quality classroom tasks that teachers could use on a voluntary basis. These classroom activities are intended to both provide educators with high-quality assessments for their local assessment systems while supporting professional learning of the NGSS. The second component of Kentucky’s system is a modest number of through-course tasks that school districts are required to administer to help calibrate local expectations for student work (KDE, 2017b). The state is not using the classroom or through-course results for state reporting or accountability. Finally, KDE is

implementing a relatively short-state summative science assessment in Grades 4, 7, and 11 to meet federal reporting and accountability purposes. Like Delaware, Kentucky’s system is designed to coherently assess the same learning targets following the same vision of science learning while potentially forgoing some technical quality on its state summative assessment by maintaining a firewall among the various components. However, Kentucky, after careful consideration of its theory of action, has focused on maximizing the utility for what it considers its most important purposes—supporting both student and educator learning of the NGSS.

Shifting from science, Wyoming and Utah have recently awarded assessment contracts requiring the development of interim assessments explicitly tied to the states’ summative assessment in reading and math. The interim assessments in Wyoming are based on a “modular” design whereby interim tests are tied to key subdomains within the standards (e.g., number-base 10). Importantly, in both cases, the use of the interim assessments is optional, and in Wyoming districts can administer the specific interim tests when they best fit within the local curriculum. While the states acknowledge that these are not fully balanced assessment systems, they are designed to eliminate some incoherence between the state summative assessment and the various district-purchased commercial interim assessments. These examples illustrate how states can support coherent assessment approaches. This is likely as far as states can go without having closer ties to the enacted curriculum. Therefore, districts remain the most fruitful organizational level for developing high-functioning assessment systems.

Discussion

These two terrific papers make a compelling case for how to move forward on more effective classroom assessment approaches. The important conceptualizations and examples presented by Wilson (2018) and Shepard et al. (2018) will be even more powerful if they are located in coherent, efficient, and useful district and perhaps state assessment systems. The science assessment systems in Kentucky and Delaware are not grounded in curricula as Shepard et al. (2018) discussed nor are they as psychometrically tight as the examples in Wilson (2018). Yet, they provide a way forward toward coherent, albeit loosely coupled, systems of assessment that are a considerable improvement over current practice.

If districts are to be the loci of control for balanced assessment systems, how can we, as a field of measurement experts, help ameliorate the considerable assessment capacity debt found in many local school districts? Shepard et al. (2018) acknowledge the challenges associated with this work: “While it is true that many individual expert teachers create amazing, caring and challenging learning environments for their students, it is not possible working alone (even with open source materials on the internet) to design coherent systems that build deep learning over time” (p. 11). The research–practice partnerships in both Wilson (2018) and Shepard et al. (2018) provide aspirational examples of what is possible, but are there incremental strategies that we can employ in the near term to move closer to this vision?

My colleagues and I are developing tools to support districts in the design of local assessment systems,¹ but such tools hinge on educators’ capacity to evaluate quality and utility. To this end, we and others have launched efforts²

to support improvements in the assessment literacy among multiple users of assessments. This is not a new effort among NCME members, but it has taken on a renewed urgency. A major challenge is helping district and state leaders discern differences in assessment quality and to learn to care about why quality matters. Several of my colleagues are piloting criteria and processes that can be used by organizations such as edreports³ to publish evaluations of commercial interim assessments and other related products.

These are incremental advances at best. A next major step toward the vision expressed by Shepard et al. (2018) might be the creation of curricular materials that teachers and others can more easily use to begin enacting the types of rich learning experiences described by Wilson (2018) and Shepard et al. (2018). This could take the form of the National Science Foundation–funded mathematics projects of 10–15 years ago⁴ or a new form. Using research-based learning progressions as the foundation for both curriculum and assessment can maximize the likelihood that both the coherence and utility criteria will be realized. Such efforts can serve as a forum for having assessment experts play an important role because the assessments included in published curriculum materials are often the weak link and would benefit from the type of measurement expertise described by Wilson (2018).

Notes

¹See the District Assessment Toolkit at www.nciea.org

²See the December 2017 NCME Newsletter and the call for assessment literacy materials (p. 9): <http://files.constantcontact.com/fd24a90e601/a943f69b-2042-426a-a241-565e46bfeff0.pdf>

³See www.edreports.org

⁴https://www.nsf.gov/pubs/2002/nsf02084/chap1_4.htm

References

- Chattergoon, R. (2016, April). *The generalized system of assessments framework: Conceptualizing systemic aspects of assessment systems*. Presentation at the annual conference of the National Council on Measurement in Education.
- Chattergoon, R., & Marion, S. F. (2016). Not as easy as it sounds: Designing a balanced assessment system. *The State Education Standard*, 16(1), 6–9.
- Hall, E. (2014). Framework to support the validation of educator evaluation systems. Retrieved November 23, 2017, from http://www.nciea.org/publication_PDFs/A%20Framework%20to%20Support%20the%20Validation%20of%20Educator%20Evaluation%20Systems_EH071114.pdf
- Kentucky Department of Education (KDE). (2017a). Science assessment system. Retrieved December 1, 2017, from <https://education.ky.gov/curriculum/conpro/science/Pages/Science-Assessment.aspx>
- Kentucky Department of Education (KDE). (2017b). Through-course tasks in the science assessment system: An opportunity for calibrating expectations. Retrieved December 1, 2017, from https://education.ky.gov/curriculum/conpro/science/Documents/Through_Course_Tasks_Article.pdf
- Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*, 9(Sup1), 188–208. <https://doi.org/10.1080/19345747.2015.1116031>
- Li, Y., Marion, S. F., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education*, 85(2), 163–185.
- Marion, S. F. (2015, September). *Coherence and comprehensiveness across political boundaries*. Presentation at the Reidy Interactive Lecture Series. Boston, MA. www.nciea.org
- Marion, S. F., & Penuel, W. R. (2017, September). *Systems of assessment to support multiple uses and users*. Presentation at the Reidy Interactive Lecture Series. Portsmouth, NH. Retrieved November 1, 2017, from https://www.nciea.org/sites/default/files/rils_presentations_2017/Systems_presentation_092917.pdf
- McCrae, A. (2017). Assessing the NGSS in Delaware. Presentation at the Reidy Interactive Lecture Series. Portsmouth, NH. Retrieved November 1, 2017, from https://www.nciea.org/sites/default/files/rils_presentations_2017/McCrae_RILs_cog_complex.pdf
- National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- National Research Council (NRC). (2014). Developing assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education. In J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty (Eds.), *Division of behavioral and social sciences and education*. Washington, DC: The National Academies 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.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34.
- Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, 37(1), 5–20.