Title

Making Measurement Important for Education: The Crucial Role of Classroom Assessment

Permalink

https://escholarship.org/uc/item/16d8q3nc

Journal

Educational Measurement Issues and Practice, 37(1)

ISSN

0731-1745

Author

Wilson, Mark

Publication Date

2018-03-01

DOI

10.1111/emip.12188

Peer reviewed

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2018, Vol. 37, No. 1, pp. 5–20

Making Measurement Important for Education: The Crucial Role of Classroom Assessment

Mark Wilson, *University of California, Berkeley, and University of Melbourne*

This article is a written version of the Presidential Address[1] I gave at the annual meeting of the National Council on Measurement in Education (NCME) in April 2017. It is a call to NCME members (and others who read this, of course) to rebalance their focus so that classroom assessments are seen as being at least as important as large-scale assessments for education (in fact, in my view, they are more important). The article reviews research literature about the effects of classroom assessment to establish its importance for education. Then, the roles of large-scale assessment are reviewed, and, in particular, it is noted how these can have negative results when the large-scale assessments are not well aligned with sound curriculum and instructional and assessment practices grounded in theories of learning. In the next two sections (a) the idea of a learning progression is described as a way to facilitate the coherence between classroom and large-scale assessment and (b) the idea of a "roadmap" is described, being the assessment components of the learning progression. This is followed by a description of an example of such a roadmap, developed for the Assessing Data Modeling and Statistical Reasoning project using the BEAR Assessment System (BAS). Finally, a concluding discussion reviews the ways that the coherence between large-scale and classroom assessments can be achieved using the BAS, and hence make measurement more important for education.

Keywords: assessment roadmap, BEAR assessment system, classroom assessment, large-scale assessment, learning progression

In this article, I give a written version of the speech I gave as the Presidential Address at the 2017 annual meeting of the National Council on Measurement in Education (NCME). The main theme of the talk is that we need to reconsider the place of classroom assessment in our thinking about educational measurement. It is well established that classroom assessment can have a strong impact on the educational success of students, whereas the effects of large-scale assessment are harder to establish. Hence, the title—I see classroom assessment as the very best way to make *measurement* truly important for education.

In the sections that follow, I will begin by giving some research background to the statement above about the importance of classroom assessment. This is followed by sections describing the roles of summative assessment, how to relate both classroom and summative assessment to the curriculum, and the role of learning progression and their associated assessment roadmaps in classroom assessment. I will use a specific example of classroom and summative assessment, the Assessing Data Modeling assessment system developed following the BEAR Assessment System (BAS), to illustrate the argument, and I finish by describing how I see the BAS as embodying as the important links between classroom and summative assessment.

The text below derives in part from my four president's messages included in the *NCME Newsletter* during my tenure (April 2016–April 2017). In addition to being the focus of my Presidential Address, this topic of classroom assessment was also a focus of the NCME 2017 Special Conference *Classroom Assessment and Large-Scale Psychometrics: The Twain Shall Meet*, held on September 12–14, 2017 at the University of Kansas in Lawrence, Kansas.

The Importance of Classroom Assessment

As professionals involved in measurement in education, we mainly focus on the development and use of various forms of tests and assessments. Much of the funding that is used to develop such tests comes, ultimately, from the decisions of policy makers in education, principally leaders of state and federal bodies charged with administering education, but also administrators at other levels of education, not to mention the many boards, committees, and organizations that also seek to influence education. This leads inevitably to the major focus of our work being from a top-down perspective, at the behest, and for the purposes of, such policy makers and administrators. This is entirely appropriate, as these are the persons and bodies that are crucial large-scale decision makers in our educational endeavors.

However, there is an alternate way to perceive the educational enterprise—from the bottom up. From this perspective, educational measurement's core activity is to help in the

Mark Wilson, University of California, Berkeley, Berkeley, CA and University of Melbourne, Melbourne, Victoria, Australia; markw@berkeley.edu.

educational progress of each student as they learn. And the agents immediately involved in that are the student and the teacher. From this perspective, the most salient moments in student learning are being orchestrated by a teacher–student pair, and that is where one might expect that the most important decisions about student learning will be made, and where the greatest impact takes place.

Before pursuing this further, it is important to define some terms. Although I have used classroom assessment above, in fact in the literature people sometimes refer to it as *formative* assessment. The definition of *formative assessment* adopted here is as follows:

> An assessment activity is formative if it can help learning by providing information to be used as **feedback**, by teachers, and by their students, in assessing themselves and each other, to modify the teaching and learning activities in which they are engaged. (Black, Wilson, & Yao, 2011)

It is important to note that it is not the material itself of the assessment that is crucial, but instead it is the use to which it is put. A similar definition can then be made for *summative assessment*:

> An assessment activity is summative insofar as it is being used to **provide a summary** of what a student knows, understands or can do, and not to help by providing feedback to modify the teaching and learning activities in which the student is engaged.

This summary could be used for decisions within the classroom, such as evaluation of a class's progress in a curriculum, individual grading, attainment of a set of standards, or for large-scale purposes beyond the classroom, such as graduation, admission to the next level of schooling, or for use in program evaluations. Again, note that it is the *use* of the assessment that is the distinguishing feature here—in fact, a specific assessment instrument could be used in both ways under different circumstances. And this is true for the location of the assessment too—assessment that occurs in the classroom, for local classroom purposes, which I refer to as *classroom assessment*, may have both formative and summative purposes. But almost all formative assessment will occur within the classroom.

The observation above about the salience of classroom assessment in education is not just an idle observation, but has been well established by a number of very broad classical research syntheses (e.g., Crooks, 1988; Natriello, 1987). An important synthesis was the seminal work by Black and Wiliam (1998), which detailed multiple types of formative assessment that have been found effective and multiple modes of feedback that affect student learning. They reported effect sizes (specifically Cohen's d) for formative assessment between .4 and .7 (Table 1). The establishment of similar results continues to this day. A recent and rigorous review by Hattie (2009) concluded that feedback was associated with an effect size of .73, while providing formative evaluation yielded .93. This range of effect sizes persists when looking into specific topic areas. For example, working in the area of writing assessment, Graham, Hebert, and Harris (2015) found that, overall, formative assessments gave an average standardized effect size of .63, with results for specific forms of feedback including an effect size of .87 for adult feedback (including teacher feedback), .58 for peer feedback, .62 for self-assessment, and .38 for computer feedback. Moreover, these values are among the highest effect sizes yet found in the education sphere (Hattie, 2009).

Table 1. Some Reported Effect Sizes

Source	Subject Matter	Effect Size
Black and Wiliam (1998)	General	4–.7
Graham, Hebert, and Harris (2015)	Writing assessment	.63
Adult (teacher) feedback	"	.87
Peer feedback	"	.58
Self-assessment	"	.62
Computer feedback	"	.38
Hattie (2009)	General	
Feedback	"	.73
Providing formative evaluation	"	.90

Clearly, then, the topic of classroom assessment is indeed crucial for the entire educational enterprise, and should be seen as the most likely pathway for educational measurement to make a positive and central contribution to education (Wilson, 1992; Wilson & Sloane, 2000).

The Roles of Large-Scale Assessment

Few members of the educational measurement community, and few members of NCME in particular, will need to have the importance of large-scale assessment pointed out to them, so I will not devote space to that here. Instead, I will start with a discussion of the multiple aspects of how large-scale assessments are deployed in the broad educational domain. In my view, these aspects split into two functions: the "information" uses of large-scale assessments and the "signification" uses.

The *information* uses of large-scale assessment are the ones that are the main focus of research, development, and application in educational measurement in general, and I would say for the majority of those who develop and evaluate measurements. By information uses I mean the many ways that the *actual results* (i.e., the information) from the measurements are used in the educational system. For example, suppose that the instrument is a State test, designed to assess students in a specific domain, such as writing. Then, the direct information uses of this test would be to provide estimates of student location on a variable of student performance in the domain of writing. The results may be used in a variety of ways: they might be aggregated across multiple levels and groupings of students, and they may also be combined with the results of other tests in various ways. Within the classroom, individual results might be used summatively by a teacher for classroom use, or for sharing with parents. They might also be used summatively in aggregation and/or combination to make educational decisions by appropriate professionals, such as teachers, parents, administrators beyond the classroom (including those at the building level, up to the State level), and educational policy makers of many kinds.

The *signification* uses of large-scale assessments are seldom referred to in research papers in educational measurement, though they are commonly understood in educational policy circles (see, e.g., National Research Council, 2001), and, in my view actually carry greater weight in the education system. These signification uses include the signaling to teachers and others of what Standards they should be teaching (i.e., because those are the Standards represented by the items in the summative tests), and of the relative weighting of those Standards (i.e., through the relative numbers of items [or scores] representing different Standards). A second use is to give teachers concrete examples of what the Standards mean through their embodiment in specific items. (A similar distinction was made in Wilson [2004], but the term "signification" is new.)

These two sets of usages are, in my view, legitimately the aim of many policy makers in calling for large-scale tests. It must be recognized that they are also somewhat limited—for instance, that the summative information provided about individual students is relatively coarse, so coarse in fact that most teachers would already know as much about their students within a month or less of starting classes.

However, these positive usages need to be seen as being balanced by complementary negative effects. In the case of information uses, for instance, there may be attempts to gain from them diagnostic information for use by teachers. This is fraught with risk, however, as there is strong temptation to try and use subscores from summative assessments at too fine a grain size compared to the actual information content of the items (or, put it the other way, without due recognition of the uncertainty of the results from using the subscores). For comments on this, see Haberman (2008, p. 14), where the problems with the subscores of specific widely used tests are described (and for later work, see Haberman & Sinharay, 2010). Equally, the (quite appropriate) qualification of summative results using such technical concepts as standard error and reliability may give policy makers undue confidence that the results of the summative assessments are "the right stuff"—this is sometimes referred to as the "white-coat" bias (i.e., because technical experts wear white coats in their labs).

And why would it not be "the right stuff" you might ask? Well, this is where the other side of signification comes through. One negative signification effect is that the summative assessments may narrow the range of the curriculum that teachers aim for by leaving out Standards that are hard to test, and hence not be included on the test—a related negative effect is through imbalances between the predominance of items relating to specific Standards and the weight of those Standards in the overall set of Standards. A second negative signification effect is that summative assessments may narrow the ways that teachers think about what a Standard means (i.e., the items may represent only the easily testable parts of specific Standards). For in-depth reviews of these issues with large-scale testing, see, for example, Herman (2008), or, more recently, Koretz (2107), and for some approaches that can help, see National Research Council (2003).

These negative effects of summative tests are compounded in a school managerial setting where the success or otherwise of teachers and/or schools is predicated principally on external summative test results, and which can have a dire effect on teacher and school morale. My colleague, Paul Black, from King's College, London has devised a visual means of expressing this. He first notes the commonly used "*CIA* –Triangle" to symbolize the relationship between *C*urriculum, *I*nstruction, and *A*ssessment, as in Figure 1. The idea here is that the purposes of education are given in the curriculum, the way it is delivered is via the instruction, and the way it is evaluated is through the assessments. Of course, one could always look in more detail—for example, it might also be the case that information could flow from assessment directly to instruction (but probably always with the background that the purpose was established in the curriculum).

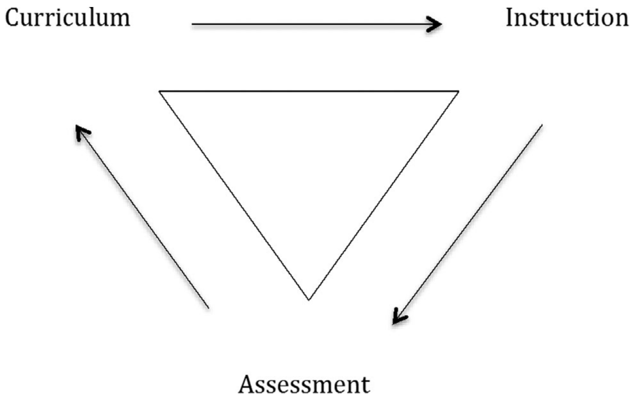


FIGURE 1. The CIA triangle.

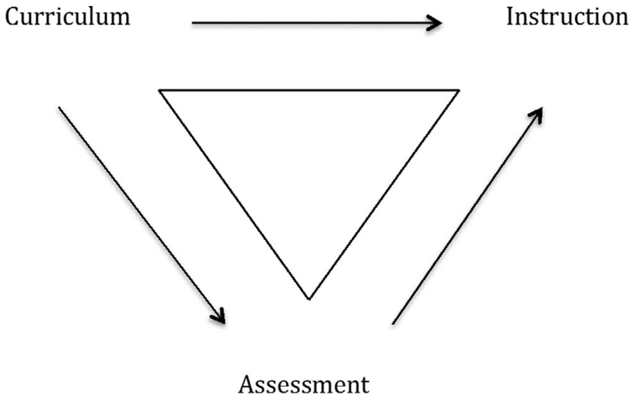


FIGURE 2. Black's "vicious" triangle.

And he then contrasts it with the situation in Figure 2, which he labels as the "vicious triangle" (Black et al., 2011) where the teachers' instructional practices are "squeezed" between the legitimate aims of the curriculum, and the focusing and narrowing effects of external (often large-scale) tests subject to the negative effects outlined above.

These negative effects of the signification uses of large-scale assessments can be hard to discern from the narrow technical point of view. Nevertheless, as I mention above, at least in my own view, the signification uses tend to be the most important in bringing about changes in the educational system (both positive and negative), and hence need to be attended to very carefully by people involved in educational measurement, as professionals, scholars, and as players in the policy realm.

Relating Classroom and Summative Assessment to the Curriculum: Learning Progressions

Black's "vicious triangle" illustrates the way that teachers' plans for their students learning can be squeezed between the demands of the curriculum and the large-scale assessments that are used for evaluative purposes. This can have multiple harmful effects, including the replacement of teaching the curriculum with "teaching the test" and related reductions in student engagement and teacher morale. The central issue is that the assessments (both classroom and large-scale) need to be working coherently with the curriculum. When that

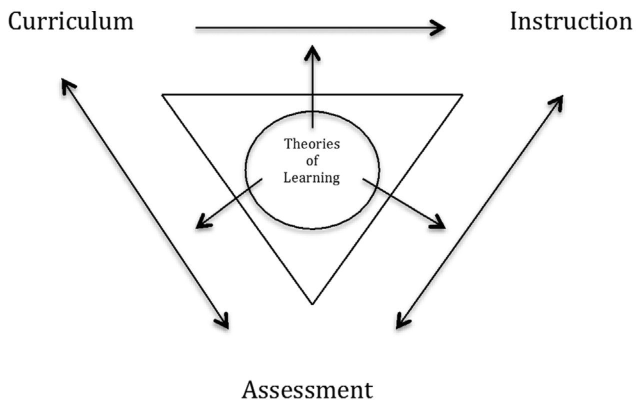


FIGURE 3. A revised CIA triangle.

coherence breaks down, or, as is sometimes the case, was never there in the first place, the sorts of negative outcomes mentioned in the previous section can be expected to occur (Wilson, 2004). Black and his colleagues have posited that what is needed are theories of learning that tie together these three elements into a coherent process (Black et al., 2011). Figure 3 illustrates this idea, emphasizing both the presence of additional forms of feedback, as well as the need for principles of coherence in the shape of theories of learning.

Thus, we must develop ways of understanding and expressing the structure of both curriculum and assessments *together* so that (i) the curriculum can be used to define the goals of the assessment (i.e., constructs to be assessed) and (ii) the results of the assessments can be interpreted directly in terms of those curriculum constructs. In my view, this is best achieved through the construction of learning progressions (also known as learning trajectories) that articulate student development through the curriculum in terms of the main content areas and the reasoning and other disciplinary practices involved. One description of the concept of a learning progression is as follows:

> Learning progressions are descriptions of the successively more sophisticated ways of thinking about an important domain of knowledge and practice that can follow one another as children learn about and investigate a topic over a broad span of time. They are crucially dependent on instructional practices if they are to occur. (Corcoran, Mosher, & Rogat, 2009, p. 37)

A learning progression should also be seen as a set of learning events for students to take part in an *epistemic culture*—an organization of social and cognitive structures as well as curriculum and assessment materials that support disciplinary modes of knowing (Knorr Cetina, 1999; Lehrer, 2009). Students must learn ways of thinking that are consistent with the standard practices of the discipline—however, at certain points toward that goal, their understanding and performances may not always be like the standard forms.

To make this development happen, progressions must not be designed from a top-down view of disciplinary progress (as would most often be seen in a typical text-book) but must instead take the shape of active experiences and events where the student interacts with the tenets of the discipline. This interaction between the student and discipline is orchestrated by the design of curriculum, instructional, and assessment materials, which constitute a theory about how to incorporate the elements of a learning ecology. This includes, for example, the types of issues, problems, and questions that are used to provoke student responses.

Learning progressions are intended as guides for teaching, and hence teachers will need to be familiar with classes of student performances that would indicate certain states in the learning progression. To do this well, it will likely require the development of new forms of assessment—these will not necessarily focus on the "right answer," but rather will need to be designed to generate performances that can reveal students' ways of thinking. To assist teachers, more will be needed than just items—learning progressions will need to incorporate descriptions of states of student learning, instructional strategies to support this type of student learning, well designed schemes of assessment that relate the states to student performances, and deep professional development that helps teachers to learn and master the new pedagogy of fostering student progress along the learning trajectory.

Learning Progressions in the Classroom

Class discussion is a central component of classroom work, as well as classroom assessment. Success for a teacher in orchestrating this depends *initially* on the power of the opening questions or activities to provoke rich discussion but, then, *second* on the capacity of the teacher to listen, to interpret the responses, and to steer the discussion in the direction of the goals of the lesson, by summarizing, or by highlighting contradictions, or by asking additional questions. To do this skillfully and productively, one essential ingredient for a teacher is to have in mind an underlying scheme of progress, or a roadmap, in the topic; such a scheme will guide the ways in which students' contributions are summarized and highlighted in the teacher's interventions and the orientation the teacher may provide by further suggestions, summaries, questions, and other activities.

This also applies to the formative role of feedback to individuals in supporting learning. Feedback, which can be verbal or written, should guide the learner, and require from the learner, to pursue further work to improve on the work already accomplished. And here, again, a clear road map is required for the teacher: (a) to formulate a task or test so that the responses can provide evidence of learning progress, (b) to formulate helpful comments, tailored to the individual needs of each student, and (c) to give clear guidance on how to improve. This road map needs to give a view of the learning aims and the steps along the route(s) that the student needs to take to get closer to that aim in light of his or her current position. Furthermore (for students of sufficient maturity), full student involvement requires that the students also have a grasp of the point they have reached along that route. The feedback must also give the student a clear aim for improvement, and if each student can locate this aim in a criterion-referenced framework, this can provide both orientation and motivation for improvement.

At the end of any learning episode, there should be review, to check before moving on, perhaps using an end-of-topic test or other forms of assessment. Here, there can be a dual purpose. One purpose is reflective; to both develop the learner's overview of the progress made and to check for gaps or misconceptions—overall, to serve as a progress review *en route* rather than as a terminal assessment. The other purpose is prospective, to look forward to building up a record of achievement, which might be a preparation for, and/or a contribution to, summative assessment.

Thus, a well thought-out and evidence-based learning progression can (i) provide the essential basis for the setting of a teacher's strategic aims, and for the planning of instruction, (ii) serve as a guide for the on-the-fly decisions that have to be taken while in the midst of teaching and assessing, and (iii) provide the criteria on which formative and summative assessment should be based. Thus, the "vicious triangle," described above, can be replaced by a better approach, where the curriculum is fashioned in terms of a model, grounded in evidence, of the paths through which learning typically proceeds as it aims for the desired targets, that is to say, the curriculum reflects and provides a strong model of progression in learning. This "road map" may then inform both pedagogy and the assessments (both formative and summative), in that an articulated set of tools can be tailored to stages in progression along the road, so that such tools will help to identify the region along the road where failure gives way to success.

To achieve this, the first issue to be addressed is to reform the interaction between curriculum and assessment, a reform that should be strongly driven by theories of student learning, but also strongly influenced by the observation and interpretation of student growth as represented in the analysis of student responses to classroom assessments. The second issue is to develop and use these learning progressions, which need to be formulated through professional judgment and the development of sound instructional practices, and confirmed by data gathered through assessments and the interpretation of student responses. For a detailed example of these steps, see Black et al. (2011).

Learning Progressions and Roadmaps

In this article, I will concentrate on the assessment aspects of this complex educational strategy. To read more about the general setting, see the example described in Lehrer, Kim, Ayers, and Wilson (2014). In Figure 4, I illustrate this complex idea of a learning progression, in a wholly inadequate way, mainly emphasizing only a few aspects of it—that it typically has a complex structure (represented by the "clouds"), that it grows in that complexity (represented by the growth in the size within the sequence of clouds), and that it is primarily an idea in the minds of the educators (represented by the small person in the bottom left-hand corner, who is thinking all of this, an educational professional).

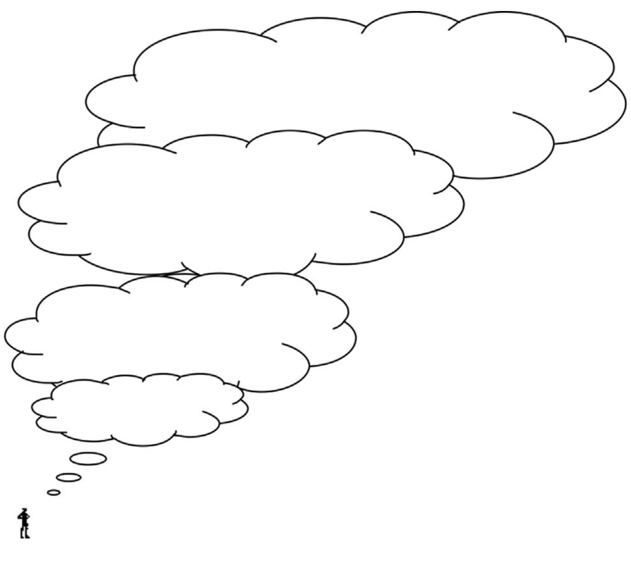


FIGURE 4. A schematic representation of a learning progression.

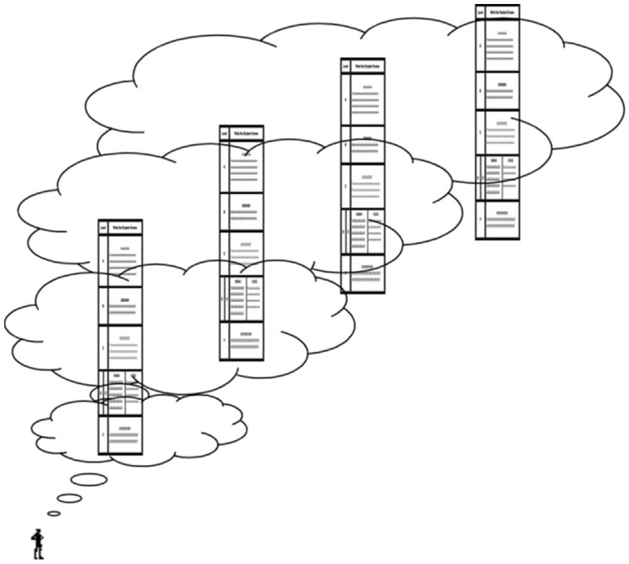


FIGURE 5. Representation of a set of constructs constituting a roadmap.

This representation can then be used as a background for the idea of "mapping" the progression in terms of, say, a set of assessment constructs, to constitute a "roadmap" that can help teachers find their way around the learning progression, and, most critically, help them track their students' progress through it. A simple version of this is shown in Figure 5. I think of this representation as being somewhat similar to lines of longitude and latitude on a globe, giving a guide to where a student is located and helping to plan the next steps. The crucial educational tactics that lead to student learning, in terms of curriculum, instructional practices, and assessments, are like the geography of the earth underlying the lines, but the cross-hatch lines are helpful in finding one's location and finding one's way. In Figure 5, the vertical bars represent the psychometric constructs designed to act like lines of longitude, mapping out the main constructs in the learning progression, and the horizontal lines in those bars are like lines of latitude, indicating important mile-posts in the roadmap of student learning (although, unlike lines of latitude, these may differ from construct to construct) In the roadmap, for each construct, there are levels that delineate different degrees of sophistication of the thinking about the construct. Note that Figure 5 is a quite simple example of a roadmap—a much more complex example is given in Figure 6. For some examples of such roadmaps, see Brown, Nagashima, Fu, Timms, and Wilson (2010—on the topic of scientific reasoning); Lehrer et al. (2014—statistics and modeling); Osborne, Henderson, MacPherson, and Yao (2016—scientific argumentation); and Wilson, Scalise, and Gochyyev (2015—ICT literacy). A very simple unidimensional example involving buoyancy can be found in Kennedy and Wilson (2007).

Armed with such a map of student development, both the curriculum developer and the assessment developer can build a coordinated system of instruction and assessment, and the resulting coherence between the two can lead to greater

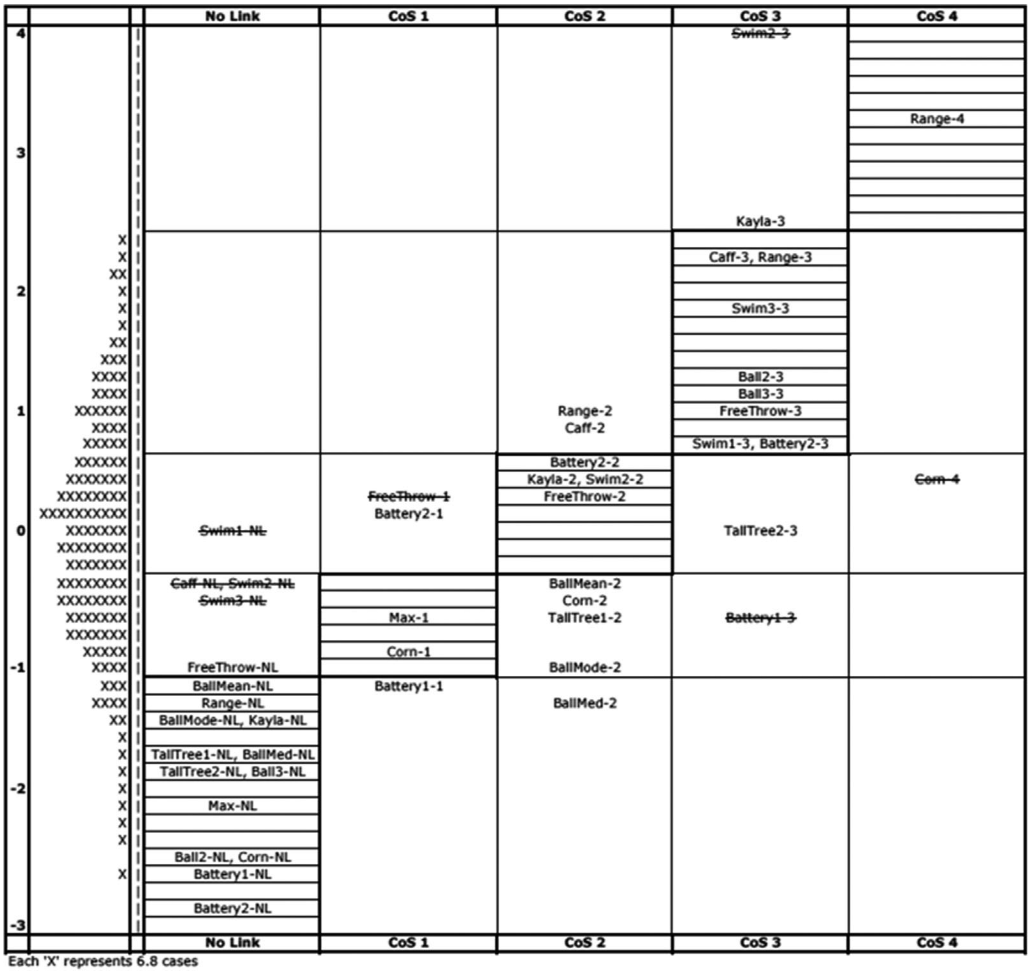


FIGURE 6. The *Conceptions of Statistics* Wright map showing cut-points.

usefulness of the assessments to instruction, and, thus, to a greater possibility of students achieving success (Wilson & Sloane, 2000). In terms of developing such a roadmap, we have found that, although curriculum ideas must be posited first, of course, it is essential that the assessment perspective be brought into consideration as early as possible, and that it is also important to include actual data from (perhaps initial versions of) assessments into the curriculum development process.

The way I create such roadmaps, they are composed as a (potentially quite complex) set of very much more simple unidimensional constructs. As an example of an empirically grounded representation of this small component of a roadmap, a unidimensional construct, see Figure 7. This "Wright map" illustrates a construct called "Consumer in Social Networks" (see Wilson et al. [2015] for more detail, where it is just one of four such constructs). The construct is shown vertically, with more sophisticated thinking toward the top. The "on-the-side" histogram of "x"s represents student locations; the numbers on the right represent score levels for items (e.g., "44.2" locates the threshold between category 1 and 2 for the three-level item 44); and the right-hand labels show the three levels: emerging consumer, conscious consumer, and discriminating consumer. This map can then be used to design assessments for both classroom assessment purposes (e.g., diagnosing individual student levels of performance, and with the augmentation of student fit statistics, to check for students with interestingly different response patterns) and summative purposes (e.g., interpreting average gains by students in different classes), and also to relate the results from these two levels together.

One concern that can be raised is that by adopting such structures for curricula, we would be constraining the choices of schools and teachers regarding their curriculum content planning. There are two points to note about this: one is that this is true in the same sense that adopting "Standards" is constraining, but indeed that is a choice most educators are comfortable with (and one might add that this constraint is somewhat stronger, due to the inherent ordering of the constructs, although such ordering is very common in Standards documents). The second is that adopting a particular structure still leaves much room for adopting and adapting a variety of instructional practices and specific educational contexts and strategies, again, just as with the adoption of "Standards." Thus, I see learning progressions as laying somewhat between curricula and standards. They could be seen as an organization of standards, which makes them more constrained than standards. But they do not necessarily make explicit decisions about instructional practices (which I see as inherent in good curricula), but certainly do constrain some instructional design issues (due to their ordering and structure).

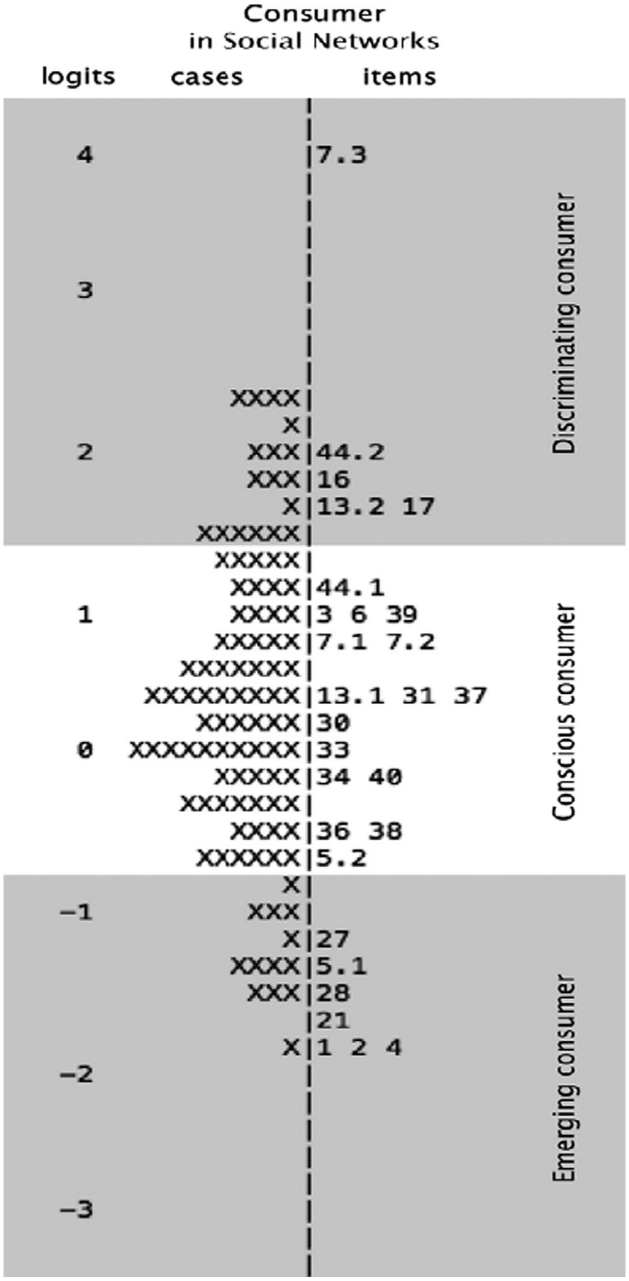


FIGURE 7. Wright map for consumer in social networks.

An Example: A Roadmap for the ADMSR Project

The Assessing Data Modeling and Statistical Reasoning (ADMSR) project was a deep collaboration between measurement and learning specialists to develop a curricular and embedded assessment system in the area of statistical reasoning for students in the middle school. It has its source in the *Data Modeling* curriculum (Lehrer et al., 2007; Lehrer et al., 2014).

Different aspects of statistical reasoning are integrated to form the *Data Modeling* approach to learning, which is illustrated in Figure 8. As the Figure shows, *Data Modeling* starts with a question about a judiciously chosen real-world phenomenon. The first step of the process is the description of important measureable attributes that are likely to inform the inquiry. Effort is then made to define and measure these attributes. The data are generated by measuring these attributes. These data must then be represented and structured in ways to facilitate the purposes of the inquiry. The statistics are employed to measure qualities of the data distributions, and probabilistic models are used to simulate inference about these statistics (Lehrer et al., 2014). An example of a teacher guide is available at the project website.[2]

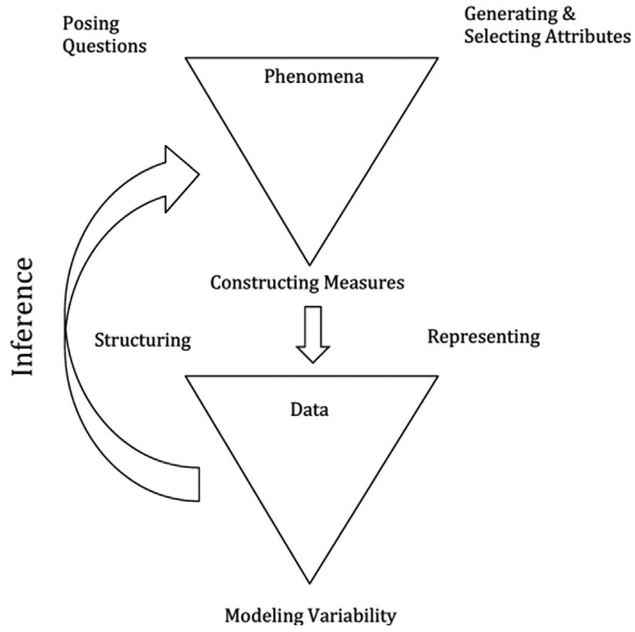


FIGURE 8. *Data Modeling* integrates inquiry, data, chance, and inference (Lehrer, Kim, Ayers, & Wilson, 2014).

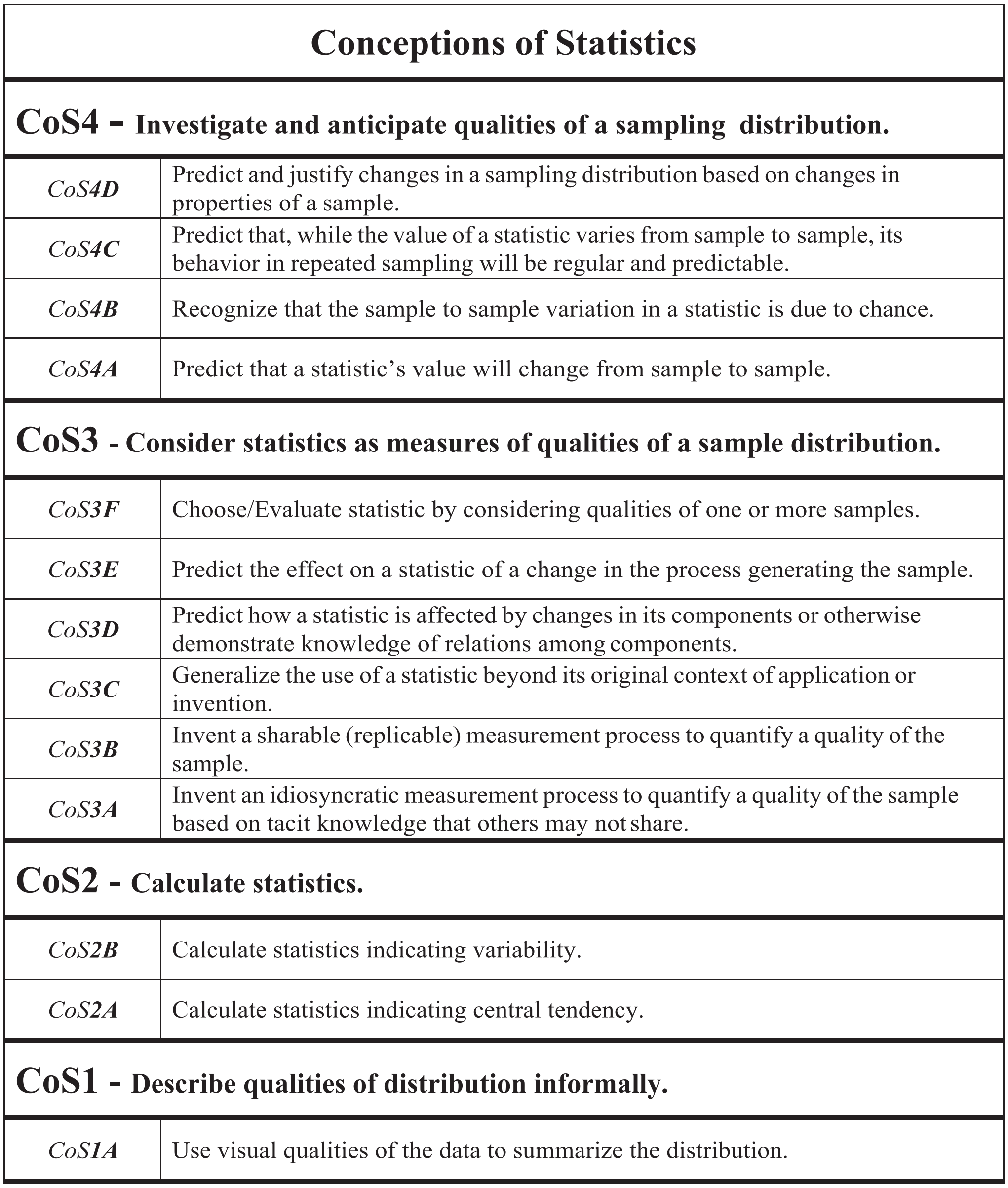
The assessments for measuring students' ability in the *Data Modeling* domains were designed and implemented following the approach in the BAS (Wilson, 2005, 2009; Wilson & Sloane, 2000). This approach to assessment is based on four principles: (1) a developmental perspective, (2) a match between instruction and assessment, (3) management by teachers to allow appropriate feedback, feed forward, and follow-up, and (4) the generation of high-quality evidence. These four principles are embedded in the BAS's "four building blocks" (Wilson, 2005):

- Construct maps
- Items design
- Outcome space
- Measurement model

In the following sections, I discuss how these building blocks have played out in the ADMSR project.

The Construct Map

The first building block, the *construct map*, is a description of a construct focusing on one major educational characteristic, and represented is an ordering of qualitatively different sign posts of performance. A construct map is used to represent one of the constructs in a cognitive theory of learning consistent with a developmental perspective, and is based on a judicious mixture of research and sound professional practice. Figure 9 shows an example of one of the construct maps from the ADMSR project, the Conceptions of Statistics (CoS) construct map (see below for more detail). The construct map is built on the idea that the construct being measured will be educationally useful if it is thought of as a continuum



Conceptions of Statistics	
CoS4 - Investigate and anticipate qualities of a sampling distribution.	
CoS4D	Predict and justify changes in a sampling distribution based on changes in properties of a sample.
CoS4C	Predict that, while the value of a statistic varies from sample to sample, its behavior in repeated sampling will be regular and predictable.
CoS4B	Recognize that the sample to sample variation in a statistic is due to chance.
CoS4A	Predict that a statistic's value will change from sample to sample.
CoS3 - Consider statistics as measures of qualities of a sample distribution.	
CoS3F	Choose/Evaluate statistic by considering qualities of one or more samples.
CoS3E	Predict the effect on a statistic of a change in the process generating the sample.
CoS3D	Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among components.
CoS3C	Generalize the use of a statistic beyond its original context of application or invention.
CoS3B	Invent a sharable (replicable) measurement process to quantify a quality of the sample.
CoS3A	Invent an idiosyncratic measurement process to quantify a quality of the sample based on tacit knowledge that others may not share.
CoS2 - Calculate statistics.	
CoS2B	Calculate statistics indicating variability.
CoS2A	Calculate statistics indicating central tendency.
CoS1 - Describe qualities of distribution informally.	
CoS1A	Use visual qualities of the data to summarize the distribution.

FIGURE 9. The *Conceptions of Statistics* (CoS) construct map.

of ability marked out by "progress points" with less sophisticated aspects of the construct at the bottom of the construct map going up toward more sophisticated points at the top end of the map. These "progress points" are also often labeled as "levels." Between each of the progress points, there are subcategories (which may or may not be ordered depending on the content). The CoS construct map displayed here will be shown in greater detail in the following. A more detailed account of the CoS construct is given in the appendix.

The participants in the ADMSR project conceived of the scientific practices of data gathering and statistical analysis as a part of modeling, and this leads the project in the direction of particular dimensions and states of learner knowledge (Lehrer et al., 2014). These dimensions and states were debated and revised through a series of design studies—these were conducted initially by the researchers and later by groups of teachers, who were not part of the original design effort, working in collaboration with the researchers. The design studies were based within school sites where we could examine relations between the instruction, assessment, and learning. To accumulate evidence about student learning, the project created assessments that could be used for both formative and summative purposes. Thus, assessment, instruction, and learning became progressively interleaved as we developed materials that made connections among professional development, learning, and assessment.

Following the procedures described in the previous paragraph, the ADMSR project created a framework of seven constructs that describe what they see as the important dimensions of statistical learning. The seven constructs (which are sometimes called "progress variables") were developed using a set of design experiments that explored typical patterns of educational growth as students began to devise and revise models of data in the *Data Modeling* curriculum. The first construct, *Theory of Measurement* (ToM), taps the sophistication with which students understand the mathematics of measurement and develop skills in measuring. This construct represents the basic area of expertise within which the rest of the constructs act. The second construct, *Data Display* (DaD), describes levels of sophistication with which students read and construct graphical representations of the data—this progresses from an early focus on cases to reasoning based on the properties of the set of data as a whole. *Meta-Representational Competence* (MRC), which is closely related, encapsulates critical modes of understanding as students learn to deploy representations for illustrating important aspects of the data and to reason about trade-offs among different representations. Another construct, CoS, is composed of a set of landmarks that students progress through from initially recognizing that statistics measure certain characteristics of the data distribution, (e.g., center and spread), to developing CoS as general tendencies, but also as subject to sample-to-sample variation (see below for more detail about CoS). *Chance* (Cha) illustrates how students' understanding about chance and elementary probability progresses as they come to appreciate that they can simulate distributions of outcomes. The *Models of Variability* (MoV) construct relates to the levels of reasoning students display when they are using probability to create a distribution of measurements. The seventh construct, *Informal Inference* (InI), illustrates how students' inferences based on single or multiple samples increase in sophistication as they go through the *Data Modeling* curriculum.

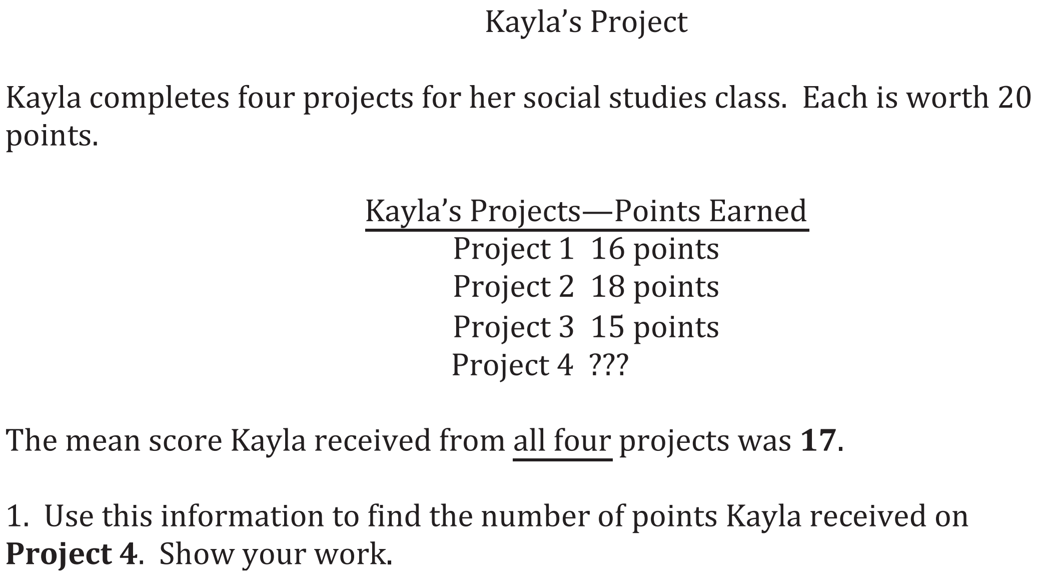
The conceptions of statistics construct. In order to better illustrate the level of detail available in the construct maps, the CoS construct is described in detail in this paragraph. As its name indicates, the CoS construct encapsulates typical student progress in understanding the concepts of statistics. It is based on the idea that statistics are summary measures of data that are designed to help examine research questions about distributions. It is crucial that students come to think of statistics as ways to summarize characteristics of sample distributions. It is important that they not see statistics as merely as an obligatory algorithmic step in data analysis. To see this, look back to Figure 9, where the levels of the CoS construct map are shown. At level **CoS1**, students tend to express the characteristics of distribution in an informal way by using the "optics" of data such as identifying clumpings, pointing out gaps, and/or mentioning the "range" of the data. At the **CoS2** level**,** students can calculate statistics, but typically do not reason about the statistic as a measure of a characteristic of a distribution. For example, a student might correctly calculate a mean but not think to relate that calculated value as a measure of the center of the distribution. The **CoS2** level has two sublevels that distinguish between statistics that are related to central tendency (**CoS2(a)**) and those that are related to variability (**CoS2(b)**). At the **CoS3** level**,** students can think about statistics as measures of characteristics of a distribution, such as center and spread. And thus, they can consider the effects of data-driven changes on the distribution, such as the influence of extreme values, and hence on a statistic (**CoS3(d)**). The first step at this level (**CoS3(a)**) begins with devising or adapting different ways to summarize characteristics of a distribution and next involves recognizing that certain statistics may be better in given specific contexts (i.e., particularly the process that is believed to have generated the distribution) (**CoS3(b)**). At the **CoS4** level**,** students initially will be noting and expecting sample-to-sample variability in a statistic (**CoS4(a)**). Next, they learn to ascribe this variability to chance (**CoS4(b)**). As students further explore sampling variability, they come to see patterns of variability that can be described by a *sampling distribution*. Eventually, students get to the point where they notice that, although changes in the location of the mean are expected from sample to sample, the variability of the sample means is lower than the variability of the values within each sample (**CoS4(c)**). This is crowned when students can predict the effect of changes in a sample on the sampling distribution (**CoS4(d)**).

The Items Design

The *items design* is the second building block of the BAS. With this building block, items are designed to provoke specific responses that provide evidence about a respondent's performance with respect to the levels of the construct map. The essential purpose of a set of items in the BAS is to be able to produce responses at all level of the construct map. The items can range over different types, whatever is best in the context. In the ADMSR project, the items included some multiple-choice items, but mostly consisted of short constructed response items. An example of the *Kayla's Project* item is shown in Figure 10. This item assesses a small aspect of student understanding on the CoS construct. We are able to evaluate their understanding of the statistical average, the mean, and especially how it is calculated from the data values.

Working with teachers to create classroom assessments. As we worked with teachers in the ADMSR project, we found we needed to transform the materials and sometimes the concepts that had been used to structure the development of the assessments. One such change involved making new versions of the construct maps to better illustrate how to use them to understand student talk and activity in classrooms, beyond the interviews and item responses we had used initially. When teachers got hold of the items, they wanted to use them to foster growth along the signpost points described in the construct maps. Hence, working with teachers we revised the depictions of the construct maps accordingly. Multimedia versions of the maps were developed. One such was a set of classroom-based, video examples of learning performances related to the levels of each construct.[3] Some early examples came from the design studies classrooms, but eventually the teachers themselves gathered vignettes from their own classes. A second was videos of *formative assessment conversations* that showed how teachers could use items to enable conceptual change. Annotations to these videos pointed out how student conversations and verbal responses, as well as student activities, pointed to particular levels of one or more constructs. These videos and the associated materials enabled teachers to view student learning performances more dynamically, as the formative assessment conversation unfolded.

The ADMSR assessment system changed as teachers engaged in two forms of assessment practice. First, teachers gave quizzes in their classrooms, based on subsets of the ADMSR item bank that the project developed. When they attended the assessment moderation meetings that were a regular part of the project, they brought along samples of their students' work that they had coded with the scoring guides developed by the project. At these meetings, the teachers became deeply engaged in discussing their own and other teacher codings of this work. This was essential for their full engagement in understanding the construct maps and the scoring guides.



Kayla's Project

Kayla completes four projects for her social studies class. Each is worth 20 points.

Kayla's Projects—Points Earned	
Project 1	16 points
Project 2	18 points
Project 3	15 points
Project 4	???

The mean score Kayla received from all four projects was **17.**

1. Use this information to find the number of points Kayla received on **Project 4**. Show your work.

FIGURE 10. The *Kayla's project* item.

Moreover, their coding sometimes resulted in modification to the scoring guides (though not generally to the construct maps).

Second, teachers also developed ways to use the items themselves as the basis for instructional materials. One way they did this was to use sample items as objects of discussion in the formative assessment conversations mentioned above. In this activity, the teachers would ask the students to solve the items individually, and then have them share their answers with the class. This would result in students seeing examples of responses at a subset of the levels of the construct map. Teachers would ask the students to decide which answer they thought was best, and have some give explanations of which were better and why. This would lead to the students themselves engaging with the progression of increasingly sophisticated levels of response that the class had generated. In adapting the project to this, we came up with construct-oriented guides for starting and structuring classroom conversations, including (a) choosing particular student responses for classroom sharing by noting their link to specific levels of the construct map, (b) contrasting different levels of reasoning by setting up (constructive) comparisons among student responses, and (c) for students whose responses were not chosen in the above, encouraging their involvement by having them choose which response was similar to their own, or alternatively having them say how hearing other responses had helped them think more broadly about their own position.

Our analysis of the classroom interactions in the project has led us to identify four forms of classroom assessment conversations. In the first one, teachers utilize the assessment items mainly as opportunities for estimating the correctness of student performances (not necessarily closely attuned to the levels of the construct map—more like a dichotomous right/wrong). These types of classroom conversations featured an "IRE" (Initiate-Respond-Evaluate) plan of classroom discourse (Mehan, 1979). The second type of formative assessment conversation was aimed at using items to generate possibilities for student participation. The classroom conversations tended toward a show-and-tell format, with students taking turns to add their pieces. We found it difficult to identify patterns for the selection and ordering of student presentations. Later follow-up with teachers generally tended to support the idea that the selection and ordering were not related to the content. A third type of assessment conversation was based on planful encouragement of level-aligned student responses to explore the local diversity of forms of reasoning that are only summarized in the construct maps and scoring guides. Typically, this was formatted as an ordered set of conversation/demonstrations, with time at the end set aside for classroom questions and follow-up (and this is what is described in the previous paragraph). The final type of assessment conversations was based on the third approach, but went beyond that by including teacher comparisons between the different levels of reasoning. Having a typology of assessment conversations meant that we could also map teacher advancement in assessment practices.

The Outcome Space

As soon as the items have been given to the students, the resulting responses can be interpreted using the third building block, the *outcome space*. The outcome space describes in detail how a student's responses to items are to be coded to different points on the construct map. Each item in the ADMSR assessment provides information about a student's level on one or more of the seven constructs. Scoring exemplars were developed which code these responses to a level on each construct. A sample of such exemplars for the Kayla's project item is shown in Figure 11.

As can be seen, level 3 is the highest level of CoS informed by this item. At that (abstract) level, the student can deploy flexible strategies for solving this problem. Specifically for this item, the student must first understand that if the mean of the four scores is 17, then the scores must add to 68, and hence, they can then find the unknown score by subtracting the sum of the given values from 68. In contrast, a student at level 2 can understand how to calculate the mean and use the formula as they usually would when provided a data set. For this item, students might use a "guess and check" strategy to attempt

Level	Response Description	Example Student Responses
3D	Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among its components	1. The differences between the mean and each score are -1, 1, -2, so the last difference must be 2 and the score must be 19. 2. 16 17 68 18 X 4 -49 +15 ------ ----- ----- 68 19 49
2A	Calculate statistics indicating central tendency.	$\begin{array}{r} 16 \\ 18 \\ 15 \\ \underline{20} \\ 68 \\ \div\ 4 \\ \hline 17 \end{array}$
NL(ii)	Student begins to carry out a strategy, but not to completion.	18 16 +15 ----- 49

FIGURE 11. Scoring exemplars for the *Kayla's project* item.

to solve the problem. Students who gave responses judged to have relevant terms in them, but that did not provide evidence of performing at a specific level on the CoS construct, were scored a "NL(ii)," while those who gave completely irrelevant responses were scored a "NL(i)." The "NL" responses are recorded in this way to capture responses to the item that have "no link" to the CoS levels. Those students who were administered the item but failed to provide a response were scored as "missing."

The Measurement Model

The *measurement model* is the final building block of the BAS. It is a 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. Different measurement models can be applied to a given instrument for different reasons. In this case, the project initially applied a unidimensional measurement model (the partial credit model, Masters (1982)) to each of the constructs to help evaluate whether the theory of development encapsulated by the levels of the construct maps is consistent with the empirical results.

The empirical data and results from initial data analyses were used to edit and/or remove inconsistent CoS items, as well as the items that had a very low numbers of responses at a given level (Schwartz, Ayers, & Wilson, 2017). Then, cut-points were set for the CoS construct and a new Wright Map with these levels is displayed in Figure 6.

The Wright Map in Figure 6 includes horizontal lines that represent the cut-points between the construct levels and shaded boxes that indicate items that behave within the cut-points for a given level. The mean thresholds for the items at different levels are mainly increasing as would be expected. The one area where overlap appears to be an issue is at the lowest threshold estimates for CoS level 2. These overlaps could indicate two possibilities: (1) the adjacent levels are not clearly distinct from one another, or (2) the overlapping items have certain features that modify their difficulty beyond the effects of the CoS levels. As the overlap in this case is limited to only a few items, Battery2 on level 1 and BallMed and BallMode on level 2, a next step would be to examine these items carefully (probably including some cognitive labs as part of the investigation) before making any conclusions about reconsidering the levels of the construct map. The results presented here for the CoS construct support the existence of the developmentally ordered levels set forth by the construct map, although further examination must be undertaken for some overlapping items. The uniformity of the threshold estimates across items provides us with confidence in the setting of cut-points and classification of students.

When each dimension had been carefully calibrated separately, a multidimensional item response model was used to estimate student locations across a seven-dimensional vector

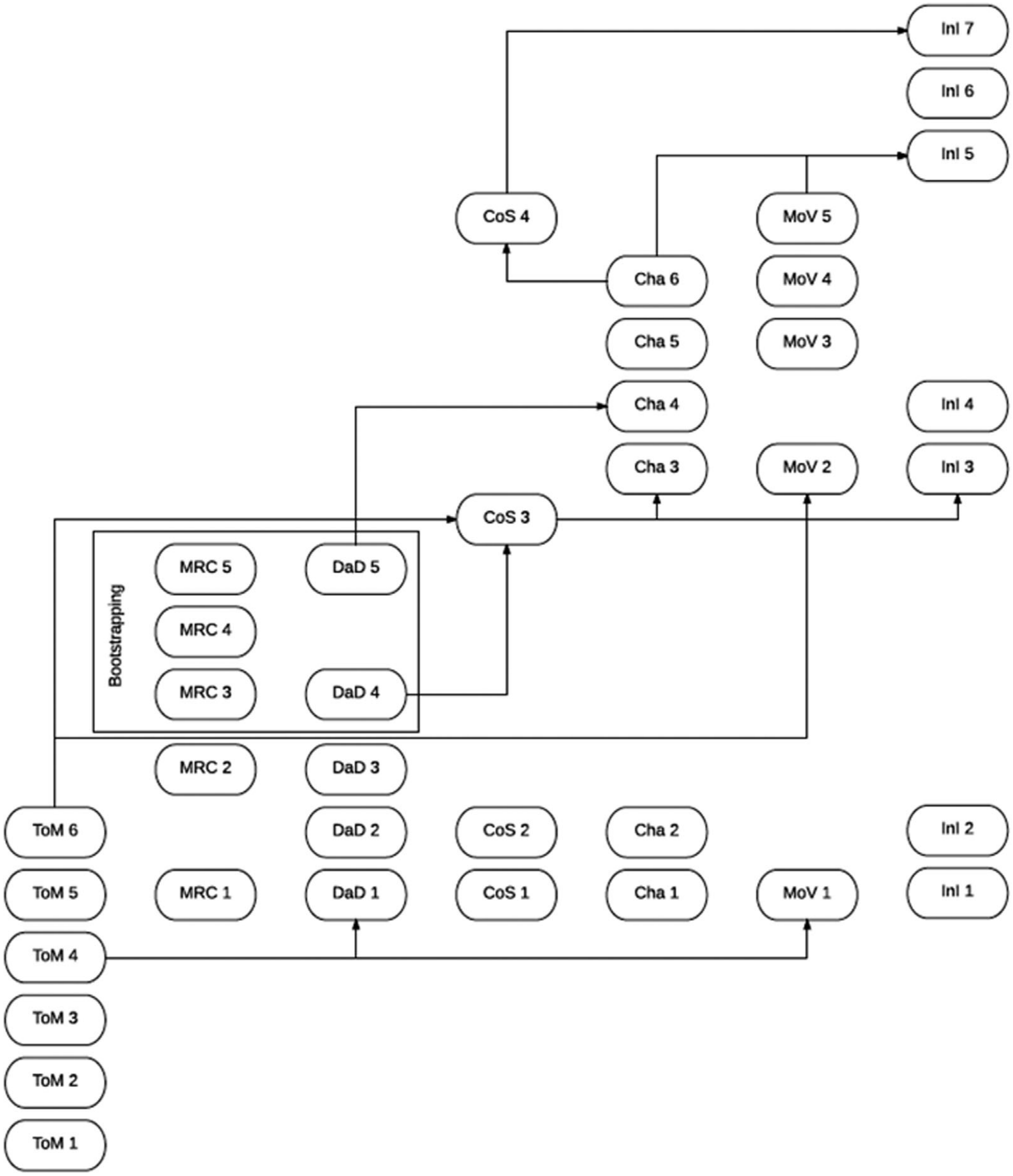


FIGURE 12. The hypothesized ADMSR roadmap.

of ability estimates (Adams, Wilson, & Wang, 1997). One form of a roadmap can be seen as being composed of a collection of construct maps like those shown in Figure 5 (Draney, 2009; Wilson, 2009), and hence, the seven constructs described above constitute a roadmap for the curriculum *Data Modeling*. However, beyond this, the ADMSR project hypothesized that a student not only advances vertically *up* a single construct map, but will also move transversally *between* different construct maps (i.e., for a student at certain levels of a construct, it will be required not only that they have mastered the previous level of that construct, but that they have also mastered a level of another construct). These hypothesized connections between the constructs are shown in Figure 12. The arrows represent specific "requirement" connections between levels of different constructs—that is, success on the level at the "point" of the arrow will not be likely unless the student is successful on the level at the base of the arrow.

For example, the MoV construct indicates a progression of understanding that culminates in modeling phenomena with chance devices. As the arrows at different MoV levels in Figure 6 illustrate, this construct relies on an orchestration of the components of the other data modeling constructs. In addition, the area noted as "Bootstrapping" between the upper levels of the DaD and MRC constructs indicates that the levels of the two constructs do not have a one-way causal connection, but rather a mutual dependence. To model, and test, hypotheses such as these requires a special form of a *multidimensional* measurement model, specifically a structured constructs model (Wilson, 2012). This is beyond the scope of this article, so I will direct the reader to Shin, Wilson, and Choi (2017) for more information.

Discussion: Relating Classroom and Summative Assessments Via the BEAR Assessment System

Following the sections above, we can see important ways that the BAS can help improve *both* summative and classroom assessments. First, the inclusion of a developmental approach in the shaping of a construct map and/or learning progressions provides an essential interpretative framing for many, if not most, uses of assessment. This makes available the resource of the *logic of development* for use in many assessment contexts, ranging from its use as a guide to moment-to-moment classroom teaching practices, to daily evaluation of student responses and student products, to the planning of educational activities over weeks and months. Beyond that, the existence of such resources provides a sound foundation for individual teacher reflection, as well as sharing between teachers, such as sharing materials and debating methods of making educational progress.

Second, by achieving a match between teaching and assessments via the items design (i.e., when the range and style of classroom activities is reflected in the range and style of assessments), then these items can help teachers design and develop new teaching activities, and can themselves be used by teachers as part of and/or prompts for teaching activities.

Third, the existence of an outcome space/scoring guide can act as a positive aspect of teacher management of both assessment and instruction in the following ways:

(a) the terms and logic of the scoring guide can establish a common language for valuing specific examples of student work/responses;

(b) example student responses included in the scoring guide can give teachers a quick, concrete, illustration of what they can expect of students at a particular level; and

(c) descriptions of the logic of scoring into levels can help teachers appreciate the sometimes subtle differences between construct map levels.

Fourth, the use of results from a measurement model can help in the following ways:

(a) when the results are visualized as Wright maps, these can be a useful metaphor for talking about student progress with students, their parents, and other teachers;

(b) the results can be used to identify, record, and track student progress and to illustrate the skills that students have mastered and those that the students are currently working on; and

(c) by placing students' performance on the continuum defined by the map, teachers and others can interpret student progress with respect to the Standards that are inherent in the constructs.

Moreover, one can see that, in an important sense, classroom assessment can, and should, be seen as the foundation for all assessments, including large-scale assessments. The BAS provides a common grounding, expressed in the construct maps, that can enable synergy between the formative and the summative, as explicated in the previous sections. In particular, classroom assessments developed using the BAS approach help large-scale assessment in the following ways: (i) the construct maps make the link to the Standards explicit and promote a developmental view of the large-scale results (i.e., improve the content validity of the large-scale assessments); (ii) when the range and style of the items from the large-scale assessments match those of the classroom assessments, the large-scale assessments are a better match to the instruction (i.e., hence they improve the instructional validity of the large-scale assessments); (iii) using the outcome space(s) for large-scale assessments means that the same interpretational frames are used for both instructional activities and large-scale assessments; and (iv) mapping student large-scale results onto the same (or similar) Wright maps as are used for the classroom assessments allows students and teachers to track their progress throughout the year.

In conclusion, we can see that there are multiple ways that soundly developed classroom assessments can be the foundation for better instruction and also for better large-scale assessments, and hence for better education for our students. And this would have the side effect of making measurement more important for education.

Appendix

Conceptions of Statistics (CoS)

This construct describes the development of concepts of statistics. It reflects the perspective that statistics are summary measures of data that are developed to answer research questions. It is important that students come to see the functions of statistics as ways to characterize qualities of the sample distributions (i.e., central tendency and spread) and not merely as an obligatory procedural step in working with data.

At level **CoS1**, students describe qualities of distribution informally by using visual qualities of data such as identifying clumps, noticing holes, or discussing the "spread" of data.

At level **CoS2**, students calculate statistics, but may fail to reason about the statistic as a measure of a quality of a distribution. For example, a student may calculate the mean but neglect to relate the mean to the center of the distribution or not consider the effects of outliers on the mean.

At level **CoS3**, students conceive of statistics as measures of qualities of a distribution, such as its center and spread. Hence, they can reason about the effects of changes in distribution, such as the presence or absence of extreme values, on the resulting value of a statistic. The initial step of this level starts with inventing or appropriating different ways to summarize qualities of distribution and then includes recognition that different statistics may be appropriate given particular contexts (i.e., the process generating the distribution) and forms of distribution.

At level **CoS4**, students begin by noting and expecting sample-to-sample variability in a statistic and attribute this variability to chance. As students investigate sampling variability, they come to understand regularities in variability that can be described by a sampling distribution. For example, students may realize that although changes in the location of the mean are expected from sample to sample, the variability of the samples' means is lower than the variability of the measurements constituting each sample. This culminates in predicting the effects of changes in properties of a sample on the sampling distribution.

Level		Performances		Examples
CoS4	**Investigate and anticipate qualities of a sampling distribution.**	*CoS4* **D**	Predict and justify changes in a sampling distribution based on changes in properties of a sample.	▪ Students predict that the variability of a sampling distribution of the median will change if the sample size is decreased from 30 to 3 and explain why (grade 5 FP study).
		CoS4 **C**	Predict that, while the value of a statistic varies from sample-to-sample, its behavior in repeated sampling will be regular and predictable.	▪ "If we measure the teacher's arm span again and again, we will get different means and spreads. However, they will not be very different next time." ▪ "If we tested another 75 Type A batteries, I would expect the median to be similar to the median we got this time and around this area (between 165 and 175)."
		CoS4 **B**	Recognize that the sample-to-sample variation in a statistic is due to chance.	▪ "From sample to sample, the medians change, just by chance."
		CoS4 **A**	Predict that a statistic's value will change from sample to sample.	▪ "If we measured the height again, I won't expect the mean to be exactly the same, even if we use the same tool and the same method."
CoS3	**Consider statistics as measures of qualities of a sample distribution.**	*CoS3* **F**	Choose/Evaluate statistic by considering qualities of one or more samples.	▪ "It is better to calculate the median because this data set has an extreme outlier. The outlier increases the mean a lot." ▪ "The estimate of the mean will be better if we can increase the number of cases, because the mean measures central tendency and more cases increases our confidence in this tendency."
		CoS3 **E**	Predict the effect on a statistic of a change in the process generating the sample.	▪ "If we use a more precise measurement tool, our spread number will get smaller." ▪ "The average deviation of rates of change of fast plants will increase as the plants grow, because of the growth spurt."

CoS3 ***(continued)***		*CoS3* **D**	Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among components.	▪ "If we increase the highest value, the mean will change, but the median will not." ▪ "If I know the mean and all but one of the data values, I can find the missing value."
	Consider statistics as measures of qualities of a sample distribution.	*CoS3* **C**	Generalize the use of a statistic beyond its original context of application or invention.	▪ "Nick's measure of spread works because when the data get more spread out, it increases." ▪ Students use average deviation from the median to explore the spread of the data across multiple samples.
		CoS3 **B**	Invent a sharable (replicable) measurement process to quantify a quality of the sample.	▪ "In order to find the best guess, I count from the lowest to the highest and from the highest to the lowest at the same time. If I have an odd total number of data, the point where the two counting methods meet will be my best guess. If I have an even total number, the average of the two last numbers of my two counting methods will be the best guess."
		CoS3 **A**	Invent an idiosyncratic measurement process to quantify a quality of the sample based on tacit knowledge that others may not share.	▪ "In order to find the best guess, I first looked at which number has more than others and I got 152 and 158 both repeated twice. I picked 158 because it looks more reasonable to me."
CoS2	**Calculate statistics.**	*CoS2* **B**	Calculate statistics indicating variability.	▪ "We found the range by subtracting the minimum value from the maximum value."
		CoS2 **A**	Calculate statistics indicating central tendency.	▪ Students calculate mean, median, and mode when they are given a set of data and put these as labels in their displays. However, they may not understand that each is a measure of central tendency.
CoS1	**Describe qualities of distribution informally.**	*CoS1* **A**	Use visual qualities of the data to summarize the distribution.	▪ "There is a big clump here." ▪ "The measurements are really spread out." ▪ "The majority is in the middle." ▪ "The real value might be where most of measurements are."

Acknowledgments

I would like to acknowledge the thoughtful comments and advice from Karen Draney and Lorrie Shepard. Any errors are, of course, my own responsibility.

Notes

[1] Available at http://www.ncme.org/ncme/NCME/News1/Presidents_Message/Past_Presidents_Speeches/NCME/News/PastPresidentSpeech.aspx?hkey=969c169a-7fe8-448d-bdec-cfbecac9efb9

[2] See: http://modelingdata.org/files/DM3_140805_lesson.pdf

[3] Please contact Rich Lehrer at the following email address for details on how to view these: RichLehrer@Vanderbilt.edu

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–147.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 71–123.

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M. J., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment*, *15*(3–4), 142–174.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). New York, NY: Center on Continuous Instructional Improvement, Teachers College–Columbia University.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438–481.

Draney, K. (2009, June). *Designing learning progressions with the BEAR assessment system*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: Meta-analysis. *Elementary School Journal*, *115*, 523–547.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Herman, J. L. (2008). Accountability and assessment in the service of learning: Is the public interest being served? In K. Ryan & L. Shepard (Eds.), *The future of test-based accountability* (pp. 211–231). New York, NY: Routledge/Lawrence Erlbaum.

Kennedy, C. A., & Wilson, M. (2007). Using progress variables to map intellectual development. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in schools: Intellectual growth and standard setting*. Maple Grove, MN: JAM Press.

Knorr Cetina, K. (1999). *How the sciences make knowledge*. Cambridge, MA: Harvard University Press.

Koretz, D. (2107). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

Lehrer, R. (2009). Designing to develop disciplinary knowledge: Modeling natural systems. *American Psychologist*, *64*, 759–771.

Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey, & K. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31–60). Charlotte, NC: Information Age.

Lehrer, R., Schauble, L., Wilson, M. R., Lucas, D. D., Karelitz, T. M., Kim, M., & Burmester, K. (2007, April). *Collaboration at the boundaries: Brokering learning and assessment improves the quality of education*. Paper presented at the American Education Research Association Annual Meeting, Chicago, IL.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mehan, H. (1979). *Learning lessons: Social organization in the classroom*. Cambridge, MA: Harvard University Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Washington, DC: National Academies Press.

National Research Council. (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment*. *Workshop Report*. Committee on Assessment in Support of Instruction and Learning. Washington, DC: National Academies Press.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, *22*(2), 155–175.

Osborne, J. F., Henderson, J. B., MacPherson, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*, 821–846.

Schwartz, R., Ayers, E., & Wilson, M. (2017). Mapping a learning progression using unidimensional and multidimensional item response models. *Journal of Applied Measurement*, *18*, 268–298.

Shin, H.-J., Wilson, M., & Choi, I.-H. (2017). Structured constructs models based on change-point analysis. *Journal of Educational Measurement*, *54*, 306–332.

Wilson, M. (1992). Educational leverage from a political necessity: Implications of new perspectives on student assessment for Chapter 1 evaluation. *Educational Evaluation and Policy Analysis*, *14*(2), 123–144.

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. 103rd yearbook of the National Society for the Study of Education, Part II. Chicago, IL: University of Chicago Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching*, *46*, 716–730.

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: Hypothesized links between dimensions of the outcome progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 317–343). Rotterdam, The Netherlands: Sense Publishers.

Wilson, M., Scalise, K., & Gochyyev, P. (2015). Rethinking ICT literacy: From computer skills to social network settings. *Thinking Skills and Creativity*, *18*, 65–80.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181–208.