



Center for
K–12 Assessment
& Performance Management

*An independent catalyst and resource for the improvement of
measurement and data systems to enhance student achievement.*

Exploratory Seminar:

Measurement Challenges Within
the Race to the Top Agenda

December 2009

Assessment for Learning and for Accountability

Mark Wilson

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Copyright © 2010 by Educational Testing Service. All rights reserved. ETS is a registered trademark of Educational Testing Service (ETS).

Assessment for Learning *and* for Accountability

Mark Wilson

University of California, Berkeley

This paper was presented by Mark Wilson at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of all of the papers presented at the seminar at <http://www.k12center.org/publications.html>.

In this paper I argue that the types of accountability systems prevailing at this time have damaged what should be the regular and appropriate connection between the curriculum (as represented by state standards) and instruction and assessment within the classroom, leading to a deformation of their characteristics and functions within the educational system. Specifically, under the auspices of state accountability systems such as No Child Left Behind (NCLB) in the United States and the National Curriculum Assessments in the United Kingdom, the format of the state standards that define the state curriculum has led to a particular type of summative assessment that has overwhelmed formative assessment, and thus taken over the commanding role in guiding classroom instruction. In this way both formative and summative assessments take on roles that in combination restrict the curriculum and hamper good classroom teaching. In contrast, I propose that there is another way that standards can be developed and communicated so that the roles of formative and summative assessment in schooling can be combined (or at the very least rendered complementary and synergistic) and improved, to aid the success of schools. The key ingredient in creating this synergy is ensuring that (a) there is a common *roadmap* or *learning progression* that ties together the standards in an educationally meaningful way, (b) the standards themselves are expressed and communicated in a richer and more educationally effective way (which are called *learning performances*), (c) both forms of assessment (i.e., summative and formative) are based on that learning progression, and (d) the assessments have been built to provide teachers and school administrators with useful information regarding students' placement along the learning progression.

First I present the typical purposes and unique features of each of the two types of assessment—formative and summative. I then present how assessment is related to curriculum and instructional practices (in the familiar Curriculum-Instruction-Assessment *triangle of learning*). This provides a context in which to reflect on the following concern: Why are formative and summative assessment not working properly in their current forms? Responding to that, I propose to shatter this harmful hold of summative assessments on instruction by situating the two forms of assessment, and also curriculum and instruction, on a foundational model of learning—that is, a learning progression. Although I readily acknowledge that learning progressions are a generic idea, having many possible manifestations, I will describe one particular approach to the construction of learning progressions, using the BEAR (Berkeley Evaluation & Assessment Research) Assessment System (BAS; Wilson, 2005). I describe the logic behind

the BAS and give an example of its use for a German test of mathematical literacy based on the Programme for International Student Assessment (PISA) mathematics framework.

Formative and Summative Assessment

There is much uncertainty about the connection between formative and summative assessment. One view is that formative means nothing more or less than numerous, but small, summative testings that are part of a regular regime of unit/class assessments followed up by scores/grades noted and given back to students. Another, polar opposite, view is that the two are so completely dissimilar in purpose that, to make sure of the validity of each, they must be completely disconnected in concept and development. Both of these views are based on distinctive understandings of the suitable interrelationships among curriculum, instruction, and assessment. In what follows I argue for a different approach.

Formative Assessment

It is important to be specific about the meaning and practice of formative assessment. The definition I will use 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, 2009)

Beyond this definition, there are three other important features of formative assessments. The first is that feedback should be three-way—from students to teacher so that the teacher can place the students' level of achievement, from teacher to students so that the teacher can help with the students' understanding and/or broaden it, and from student to student so that students can assist one another through classroom interactions. The second important feature is that the formative assessment itself can assume many curricular grain sizes, ranging from assessments that are narrowly diagnostic in their nature to assessments (providing detailed information about the infrastructure of a skill or domain) to those that are broader in their nature (usually assessing how students orchestrate many features of the skill or domain) and, as such provide "curricularly diagnostic" information to a teacher, such as how to plan the next few lessons. The third important feature of formative assessments is that feedback can be carried out through both verbal and written exchanges, and over various time periods.

Summative Assessment

It is necessary at this point to clarify the distinction between formative and summative, assessment, starting with a definition of summative:

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. (Black & Wilson, 2009)

Summative assessments have multiple roles. In terms of the student they can be used to examine general improvement and so can provide guidance, recognition, or motivation in terms of what has been achieved; they can also constitute admission to a further level of education or employment. For the teachers of the students they can provide a global, and external, guide to the success of their efforts, and for a school or school district (or state) they can be used as a basis for measuring the accountability of the teacher and/or school. Clearly many of the state tests used in conjunction with NCLB have this summative character.

Curriculum, Instruction, and Assessment: How They Relate

The account of formative assessment above indicates that, far from being a minor add-on to an accountability system, it is central to the true purpose of the student’s learning and hence central to the purpose of the schools. Thus it should not be surprising that the way formative assessment is carried out has strong consequences for both curriculum and instruction and, moreover, that this observation implies a challenge to the conventional interpretation of the learning triangle (as shown in Figure 1).

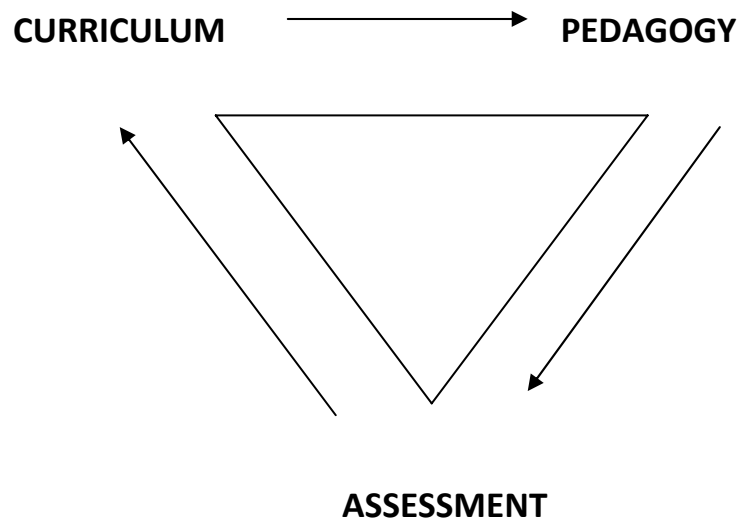


Figure 1. Traditional triangle of learning.

This traditional view starts with a limited idea of the way in which the curriculum might inform instruction. In this view the curriculum, whether prescribed by state, national, or local initiatives, or perhaps implicit in the design of school textbooks, can be seen mainly as a list of things to be taught, and the teacher’s role is seen as delivering the contents of this list through the instruction that he or she provides. Often this will be implemented by rote or recitation in which questions are simply designed to test knowledge—students compete to produce the “right answer.” In this view, which I

could label as *micro-summative*, the role of assessment is to check up on the learning, leading to grading or ranking the students, and to rewards for the successful and to exhortations for others to work harder. In this view assessment becomes the governor of the curriculum, an interpretation in which the curriculum is seen as a set of things to be known, with assessment recording which of these are known and which are not.

Unfortunately this brief sketch reflects a point of view that can be observed in many of today's classrooms, boardrooms, and statehouses. If change is to be brought about, we must understand how such a situation has arisen. One source of trouble with the operation of the standard triangle of learning is that both assessment and instruction typically have to work with a weak model of the curriculum. When the curriculum is little more than a catalogue of desirable outcomes (as it commonly is with standards-based approaches), it makes it difficult (a) for instruction to be designed in ways that promote student understanding and (b) for assessments to be designed to generate and support the interpretation of responses that lie between wholly correct and entirely wrong (which usually constitute the majority of student responses). Where assessment systems are designed to reflect a weak model of the curriculum without regard for the instruction (as will typically be the case in a high-stakes accountability environment) and, in addition, assessment tools are weak, then the assessment that is produced will lack validity, in that it will not match to any model of learning more subtle than a dichotomy between *got it right* and *got it wrong*. Then, as shown in Figure 2, the arrows in the standard learning triangle from Figure 1 can change direction, with the influence flowing to instruction from assessment, rather than vice-versa. In this situation teachers have to reconcile this influence with that of their own interpretation of the curriculum. The outcome of this *vicious triangle* for instruction is then helpful to teachers in diagnosing intermediate stages in the progress of pupils towards full understanding of the concepts or mastery of the skills in the curriculum. Then, within such an accountability system the curriculum comes to be defined (in practice) by the summative test. The teacher is consequently forced to ensure success by teaching to the test, since teaching for understanding may yield little if any reward when the test is not sensitive to the understanding (as is most commonly the case with today's tests).

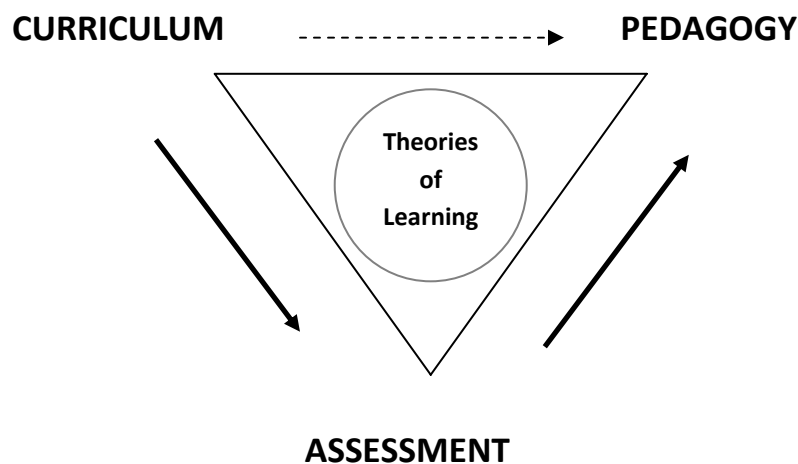


Figure 2. The breakdown of the standard interpretation of the learning triangle—the vicious triangle.

I propose to replace the resulting vicious triangle with a better approach. In this better approach a first step is that the curriculum must be designed 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 learning progression may then be used as a basis for both instruction and assessments (including formative *and* summative). A calibrated set of tools will be developed to assess landmarks along the learning progression, and these tools will be used to identify the places where failure turns into success. The validity of the assessment is enhanced because it is seen as an indicator of a student’s progression in learning. Thus the first step is an interaction between curriculum and assessment, a step that must be strongly influenced by theories of student learning, as shown in Figure 3—but it must also be strongly influenced by the observation and interpretation of student growth as represented in the analysis of student responses to assessments.

The second step is to use these learning progressions as guides to the instruction, based on the research literature and best professional practices and strongly influenced by qualitative and quantitative data from students’ responses. An important step here is the development of useful ways to help teachers and administrators (and policymakers too) interpret and use the assessment information from this new sort of assessment system. As guides they will not only help instructional planning, on both macro and micro scales, but they should also help day-to-day and minute-to-minute classroom teaching, because teachers with deep and experience-based knowledge of the learning progressions can use that knowledge in multiple ways. In this formative approach to instruction, theories of learning have a crucially strong influence, although what will then be drawn upon will include learning through discussion and group collaboration and through the development of meta-cognition.

In developing a set of learning practices (e.g., combinations of curriculum, instructional, and assessment practices) based on the formative learning triangle shown in Figure 3, curriculum and assessment developers will need to engage in numerous studies of the success of those practices. In these studies the results of the instruction will be revealed via the assessment and can then be fed back into the curriculum. Thus the arrowheads in the formative triangle point both ways, from instruction to curriculum and back, via assessment.

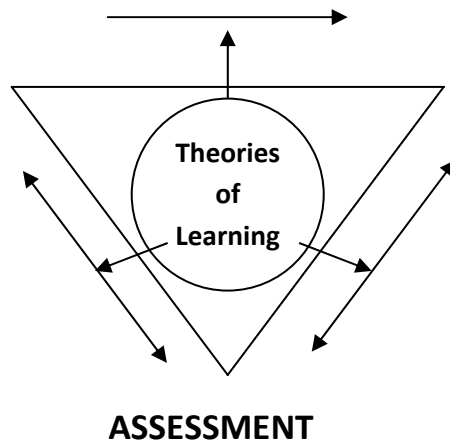


Figure 3. A formative approach to the learning triangle.

The interaction, as shown in Figure 3, does not of course resolve all assessment problems. Issues such as assessment reliability are not addressed. This is not a the same problem for formative assessment as for summative assessment, because in any given step in progression far more evidence is collected than could be elicited in a summative test, and the consequences of wrong interpretations can be quickly revealed so that such problems can be corrected. Of course reliability is still a serious problem for summative assessment; hence the design of summative and formative assessments, although based on the same learning progressions, will likely have significant differences—some in conceptualization and others in the nature and processes of development.

In terms of validity evidence, tools that closely follow the characteristics of the learning progression will most likely enhance the validity of summative judgments, but only if the nature and range of these tools is adequate. Thus a (relatively) short formal test employing only multiple-choice questions might, from a technical perspective, be better constructed than a constructed response task but might still be inadequate. In particular, in a formative setting, where a learner may have made progress to the stage of a useful but flawed understanding of an idea, it will be easier to see and appraise this in a constructed response question than in, say, a multiple-choice item.

Developmental Coherence: Learning Progressions

The argument being made in this proposal is not against the use of tests for monitoring—it is instead directed at solving the problem of making the results from the assessments interpretable and not educationally harmful. The significance of this argument goes well beyond assessments, however, as the assessment problem really arises from a curriculum problem. This concern has been brought to national and international attention by William Schmidt and his colleagues who, in their analyses of many curricula from across the world, have developed a disconcertingly appropriate description for U.S. curricula: *a mile wide and an inch deep* (Schmidt, McKnight, & Raizen, 1997). They found that, compared to the curricula in many other countries (specifically mathematics and science curricula), U.S. curricula do not develop deep understanding of subject matter, but instead tend to spread their attention across a very broad set of domains, in order to accommodate pressures from as many professional and political groups as possible. Thus typical standardized tests reflect this curricular demand (as they must, in order to survive in the marketplace). This then brings us to the topic of developmental coherence—the coherence of assessments with the patterns of development of students as they progress from novices to experts in a particular topic. I will advance a position regarding developmental coherence by focusing on the idea of a *learning progression*, and make that more concrete using the structure of the BAS.

In a recent review of the current state of research on the topic of learning progressions, the following broad description of learning progressions was given:

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.
(Center for Continuous Instructional Improvement [CCII], 2009)

The description is deliberately encompassing, allowing a wide possibility of usage, but at the same time it is intended to reserve the term to mean something more than just an ordered set of ideas or curriculum pieces. So, too, the group saw it as a *requirement* that the learning progression should indeed describe the progress through a series of levels of sophistication in the student’s thinking.

Although the idea of a learning progression has links to many older and venerable ideas in education (e.g., Bloom’s taxonomy, Bloom, 1956; Guttman scaling, Guttman, 1944; Thurstone scaling, Thurstone, 1925), the history of the specific term learning progression in the context of education is a relatively brief one (CCII, 2009), including a National Research Council (NRC) report (Wilson & Bertenthal, 2006). That report focused on assessment in K–12 education, and hence the connections to assessment have been there right from the start. A second NRC report (Duschl, Schweingruber, & Shouse, 2007) also featured the concept and enlarged upon classroom applications. Several assessment initiatives and perspectives are discussed in these reports, including references to the seminal 2001 NRC report *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001). Among the assessment programs highlighted there, probably the most prominent is the work on *progress variables* by the Australian researcher Geoff Masters and his colleagues (e.g., Masters, Adams, & Wilson, 1990; Masters & Forster, 1996), and the closely related work on the somewhat more elaborated BAS (Wilson, 2005; Wilson & Sloane, 2000). In this paper, I will draw on the latter as the core set of assessment perspectives and practices to relate to learning progressions.

Horizontal Coherence: Learning Performances

Horizontal coherence is the consistency between assessment and the other two concepts in the triangle of learning discussed above—curriculum and instruction. A concept that can be useful in considering the horizontal coherence among the three elements of the learning triangle is that of *learning performances*, a term adopted by a number of researchers—Reiser (2002) and Perkins (1998), as well as the NRC report *Taking Science to School* (Duschl et al., 2007). The idea is to provide a way of clarifying what is meant by a standard by describing links between the knowledge represented in the standards and what can be observed and hence assessed. Learning performances are a way of enlarging on the content standards by spelling out what one should be able to do when one masters that standard. For example, within science, learning performances lay out ways that students should be able to describe phenomena, use models to explain patterns in data, construct scientific explanations, or test hypotheses. For instance, Smith, Wisner, Anderson, and Krajcik (2006) summarized a set of observable performances that could provide indicators of understanding in science.

As a concrete example, take the following standard that is adapted from *Benchmarks for Science Literacy* about differential survival: “[The student will understand that] Individual organisms with certain traits are more likely than others to survive and have offspring” (American Association for the Advancement of Science, 1993, p. 124). The standard refers to one of the major processes of evolution, the idea of survival of the fittest. But it does not identify which skills and knowledge might be called for in working to attain it. In contrast, Reiser, Krajcik, Moje, and Marx (2003) amplified this single standard into three related learning performances:

1. Students *identify and represent mathematically* the variation on a trait in a population.
2. Students *hypothesize* the function a trait may serve and *explain* how some variations of the trait are advantageous in the environment.
3. Students *predict, using evidence*, how the variation on the trait will affect the likelihood that individuals in the population will survive an environmental stress.

Reiser et al. (2003) advanced the claim that this extension of the standard is more useful because it delineates the skills and knowledge that students need to master the standard and therefore better identifies the construct (or learning progression) of which the standard is a part. For example, by detailing that students are expected to characterize variation mathematically, the extension makes clear the importance of specific mathematical concepts such as distribution. Without this extension, the requirement for this important detail may have not been clear to a test developer and hence would likely have been left out of the test.

In the context of the BAS (see The Berkeley Evaluation & Assessment Research Assessment System section below), this horizontal coherence arises in deciding what to do once the student responses have been mapped onto the levels of the construct, either (depending on the circumstances) for individuals or for the group. Tools have been developed that (a) make the mapping more concrete by, for example, by providing materials such as videotapes of classroom lessons and interviews where students at those levels are made visible, and (b) show teachers options for what they might include in their planning, by, for example, linking the levels to specific lesson plans within the curriculum and to lesson videos (Wilson, Scalise, Galpern, & Lin, 2009).

The Berkeley Evaluation & Assessment Research Assessment System

The NRC (Pellegrino et al., 2001) suggested that good assessment needs to address the three inextricably linked parts of their assessment triangle as shown in Figure 4. To address these components the BAS employs four principles similar to those outlined by the NRC: (a) a developmental perspective on learning; (b) a tight link between instruction and assessment; (c) management by instructors to allow appropriate feedback, feed forward, and following up; and (d) the generation of quality evidence to make inferences (Pellegrino et al., 2001; Wilson, 2005). Note that the BAS splits the third vertex of the NRC triangle, interpretation, into two principles.

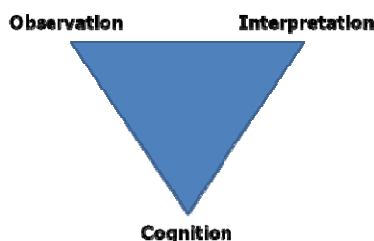


Figure 4. The National Research Council's assessment triangle.

In the BAS these four principles are expressed as the four building blocks shown in Figure 5, which are used to create an assessment: a construct map, an items design, an outcome space, and a model for making inferences about student performance (what is called a *measurement model*). The process is iterative, meaning that one is likely to move through all four building blocks several times in the process of designing an assessment.

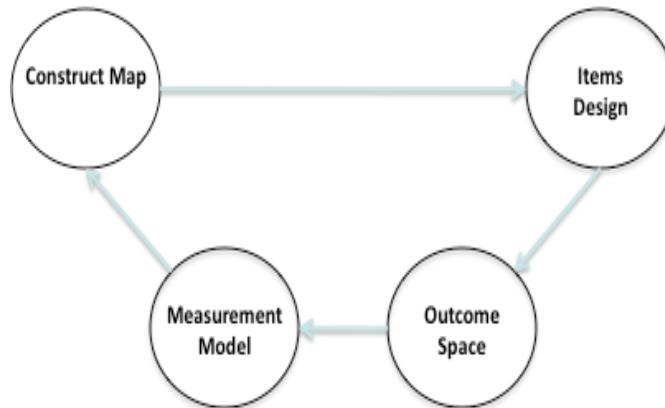


Figure 5. The Berkeley Evaluation & Assessment Research (BEAR) Assessment system.

The result is a comprehensive, integrated system for assessing, interpreting, and monitoring student performance. The BAS provides a set of tools to assess performance on central concepts in a curriculum, set standards, track progress over time, and provide feedback on student progress as well as on the effectiveness of a curriculum (Wilson, 2005).

The first step is to identify the construct(s) that are to be assessed, such as mathematics knowledge. A *construct map* helps us outline the important levels in a continuum of developmental learning in the subject area. The construct map is based on a developmental perspective of student learning: Learning is conceptualized not as an acquisition of more knowledge, but as progress within a subject area (Pellegrino et al., 2001).

The example used here is from a test of mathematics competency taken from one booklet of a German mathematical literacy test administered to a random subsample of the German PISA sample of 15-year-old students in the 2003 administration (German PISA Consortium, 2004). The test was developed under the same general guidelines as the PISA mathematics test, where Mathematical Literacy is a described variable with several successive levels of sophistication in performing mathematical tasks (Organisation for Economic Co-operation and Development [OECD], 2005a).

The levels were derived from a multistep process (OECD, 2005b) as follows:

1. Mathematics curriculum experts identified possible subscales in the domain of mathematics.
2. PISA items were mapped onto each subscale.

3. A skills audit of each item in each subscale was carried out on the basis of a detailed expert analysis.
4. Field test data were analyzed to yield item locations.
5. The information from the two previous steps was combined.

In this last step, the ordering of the items was linked with the descriptions of associated knowledge and skills, giving a hierarchy of knowledge and skills that defined the progress variable. This results in natural clusters of skills, which provides a basis for understanding and describing the progress variable.

The *items design* is a framework for designing the tasks and questions that will elicit specific kinds of evidence about student learning. The most fundamental element of this design is that the responses to the item can be mapped into the levels of the construct map. One aims to create items that tap into all levels of student knowledge. This is a basic tenet of content validity outlined in *Standards for Psychological and Educational Testing* (American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education [APA, AERA, & NCME], 1999)—that the items on a test be sampled appropriately from a wide range of student abilities. Traditional testing practices (in tests used to help make high-stakes decisions as well as standardized tests) have long been criticized for oversampling items that assess only basic knowledge and ignoring more complex levels of understanding. Matching items with the construct map ensures that one will not fall into that trap and also ensures that what is assessed is what is being taught in the curriculum. (See Figure 5.)

The *outcome space* represents in detail the qualitatively different kinds of student responses that are elicited by the items. The outcome space is influenced by the form of the student responses, that is, multiple choice, short answer, or essay. But it also represents the role of the teacher/scorer in interpreting the responses and thus evaluating student learning. Central to the BAS is the creation of coding guides that can be made concrete by including examples of scored student work. This helps teachers see progress in action and more deeply understand how to tailor their instruction accordingly. It is important to note, however, that the construct map, items design, and outcome space are not static; they may change once empirical evidence (both qualitative and quantitative) is collected. Our theory of student learning must be tested, fleshed out, and revised in response to data.

Returning to the German mathematical literacy test example, the test booklet contained 64 dichotomous items; 18 of these items were selected for this example. Example items are shown in Figures 6 to 8. Each item was constructed according to topic areas and the types of mathematical modeling required. The topic areas were arithmetic, algebra and geometry. The modeling types were technical processing, numerical modeling, and abstract modeling. The technical processing dimension requires students to carry out operations that have been rehearsed, such as computing numerical results using standard procedures; see, for example, the item shown in Figure 6. Numerical modeling requires the students to construct solutions for problems with given numbers in one or more steps; see the item shown in Figure 7. In contrast, abstract modeling requires students to formulate rules in a more general way, for example by giving an equation or by describing a general solution in some way; see the item shown in Figure 8. Because a factorial design is built into the items, the responses may also be

considered data from a psychological experiment. The experimental design has two factors, topic area and modeling type. In sum, the test is a 3 X 3 design with two observations of each pair of conditions, resulting in 18 items in total.

Function

The function given by the equation $y = 2x - 1$ shall be analyzed.

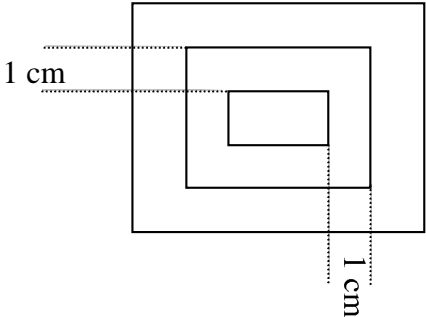
a) Fill in the missing values.

x	-2	-1	0		3	...
y					...	19

Figure 6. A technical processing item in algebra.

Rectangles

Around a small rectangle a second one is drawn. A third rectangle is drawn around the two and so on. The distance between the rectangles is always 1 cm.



By how much do the length, width, and the perimeter increase from rectangle to rectangle?

Length increases by: cm

Width increases by: cm

Perimeter increases by: cm

Figure 7. A numerical modeling item in algebra.

Difference

Put the digits 3, 6, 1, 9, 4, 7 in the boxes so that the difference between the two three-digit numbers is maximized. (Each digit may be used only once.)

1. number:	<input type="text"/>	<input type="text"/>	<input type="text"/>
2. number:	<input type="text"/>	<input type="text"/>	<input type="text"/>
	<input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 8. An abstract modeling item in arithmetic.

In the BAS, many types of items are used (e.g., open-ended, traditional multiple-choice, ordered multiple-choice) to elicit student responses. The design of these items is important, but most essential is the connection between the responses and how they are interpreted in light of the construct map.

Finally, the *measurement model* defines how inferences about student understandings are to be drawn from the scores. Here issues of technical quality are addressed. It is essential that assessments meet standards of validity, reliability, and fairness (such as consistency and lack of bias). For example, using open-ended scoring guides requires procedures for gathering, managing, and scoring student work. Raters must score the work, and issues of time, cost, fairness, and consistency arise. BEAR conducts pilot studies, trains raters, and examines reliability statistics to ensure accurate estimates of student knowledge (Brown, Dray, Lee, & Wilson, 2008; Walker et al., 2009; Wilson & Hoskens, 2001). BEAR studies the use of generalized forms of item response models for our research studies and chooses measurement models appropriately (De Boeck & Wilson, 2004; Pellegrino et al., 2001). The principal output from these measurement models is in the form of what is called a *Wright map*.

Wright maps can be very useful in large-scale assessments, providing information that is not readily available through numerical score averages and other traditional summary information. They are used extensively, for example, in reporting on the PISA assessments (OECD, 2005a). A Wright map illustrating the estimates for the Rasch model is shown in Figure 9. On this map an X represents a student. The logits (on the left-hand side) are the units of the Wright map—they are related to the probability of a student succeeding at an item, and are specifically the log of the odds of that occurring. The symbols T, N, and A represent a technical processing, numerical modeling, and abstract modeling item, respectively, with the topic area indicated by the column headings at the top of Figure 9. Where a student is located vertically near an item, this indicates that there is approximately a 50% chance of the student getting the item correct. Where the student is above the item, the chance is greater than 50%, and the further it is above, the greater the chance. Where the student is lower than the item, the chance is less than 50%, and the further it is below, the lesser the chance.

Thus Figure 9 illustrates the description of the Mathematical Literacy variable in terms of the levels defined in OECD (2005a) as well as the topic areas and the modeling types in the items design. The topic areas reflect the earlier placement of arithmetic in the curriculum than geometry and algebra. The ordering of modeling types is generally consistent with what one might expect from the definitions of the levels, except for the arithmetic abstract modeling items, which seem to be somewhat easier than expected. This is a finding that deserves a follow-up investigation.

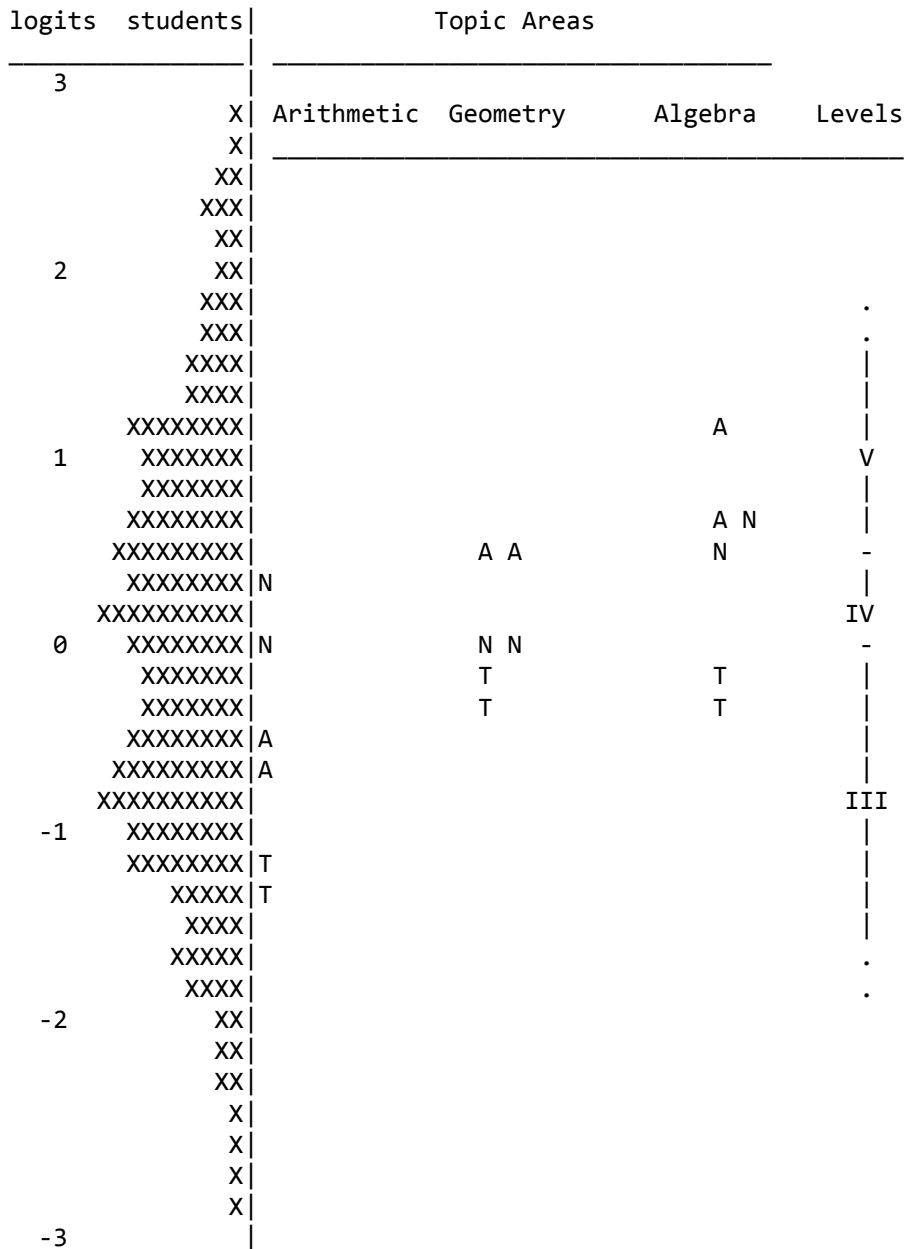


Figure 9. A Wright map of the Mathematical Literacy variable. Each X represents one case. T = technical processing, N = numerical modeling, A = abstract modeling.

Standards for Educational and Psychological Testing (APA, AERA, & NCME, 1999) describes different sources of validity evidence that need to be integrated to form a coherent validity argument. These include (a) evidence based on test content, (b) response processes, (b) test structures, (d) relations to other variables, and (c) testing consequences. In any assessment, be it classroom or large-scale, a necessary element of content validity is identifying what student progression looks like within a curriculum or a subject area and how that learning is expected to unfold. In the BAS, the first three elements of validity are accumulated through the careful development and revision of a construct map while considering test content, interview and clinical trial data incorporating student responses, and the match between the construct map and empirical item difficulties.

Last but not least, the output from these models can be used to obtain student and school locations on the construct map that can be interpreted substantively. This is necessary to ensure the assessment is useful for instruction (instructional or consequential validity). BEAR has gathered evidence for the usefulness of this approach by actively working with school districts and teachers (Brown et al., 2008; Roberts & Sipusic, 1999; Wilson & Sloane, 2000) and has created software to facilitate the use of the BAS (Kennedy, Wilson, Draney, & Tutunciyar, 2007; Wilson et al., 2009)

Mapping a Learning Progression Using Construct Maps

This section concentrates on just the first of the building blocks described above—the construct map—and potential relationships of the construct maps with the idea of a learning progression, also described above.

In order to illustrate certain aspects of the relationship between learning progressions and assessment, a visual metaphor is used that superimposes images of construct maps on an image of a learning progression. This image of the learning progression is shown in Figure 10, where the successive layers of the “thought clouds” are intended to represent the successive layers of sophistication of the student’s thinking, and the increase in the cloud’s size is intended to indicate that the thoughts become more sophisticated later in the sequence (e.g., they have wider applicability later in the sequence). The person in the picture is someone (e.g., a mathematics educator, a mathematics education researcher, an assessment developer) who is thinking about student thinking (i.e., in the example above, as students are learning mathematics).

The relationship between the construct maps that make up the learning progression may be quite complex. (See Wilson, 2009a, 2009b for examples of other relationships between the construct maps and the learning progression.) One straightforward way to see the relationship between the construct map and learning progression is to think of the learning progression as composed of a set of construct maps, each comprising a dimension of the learning progression, where the levels of the construct maps relate (in some way) to the levels of the learning progression. Note that the psychometric view of these dimensions would likely be that they are positively correlated, and hence might be illustrated as dimensions in three-dimensional space.

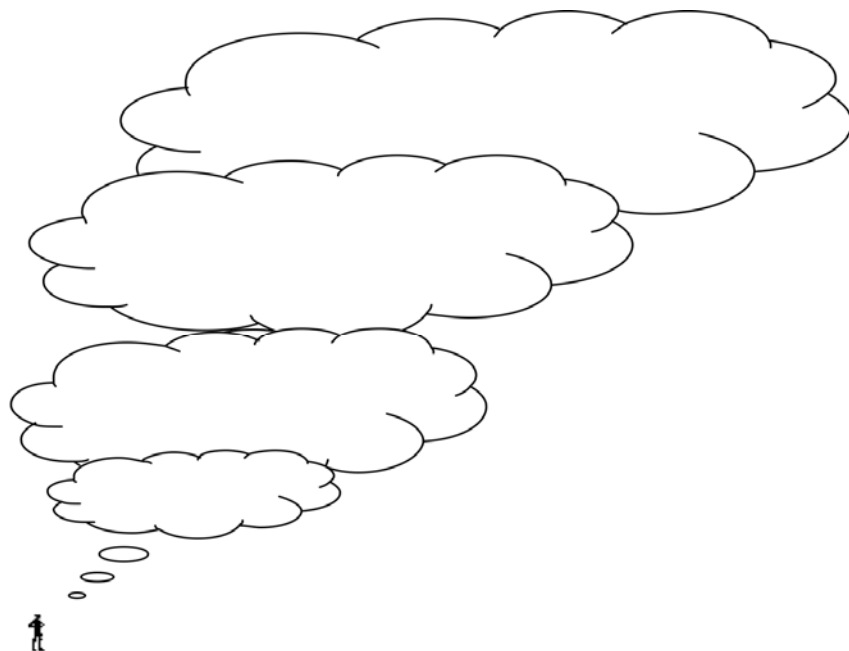


Figure 10. An image of a learning progression.

To illustrate this assessment structure, I will use a much-reduced illustration of a construct map, which will be used as an icon in later figures to represent a specific (but generic) construct map. This icon is then used several times in Figure 11, superimposed upon the earlier image of a learning progression, to illustrate the idea that the learning progression could be mapped out by a (small) set of construct maps. In this illustration the levels of the construct maps all align; that may indeed be the case conceptually, but it need not be required, as they might vary between construct maps. But the important point is that the *levels* of the learning progression relate to the *levels* of the construct maps.

In a second case, there could be an assumption that certain of the constructs were necessary for another. This could be illustrated as in Figure 12. Here the attainment of levels of a construct would be seen as being dependent upon the attainment of high levels of specific *precursor* constructs. An example of such thinking, this time in the case of the molecular theory of matter for the middle school level, under development with Paul Black of King's College, London (Black & Wilson, 2009), is shown in Figure 13. In this example, each of the boxes can be thought of as a construct map, but the relationship between them is left unspecified in this diagram. In particular, the density and measurement and data handling constructs are seen as providing important resources to the main series of constructs, which is composed of the other four constructs: properties of objects, properties of atoms and molecules, conservation and change, and molecular theory of macro properties.

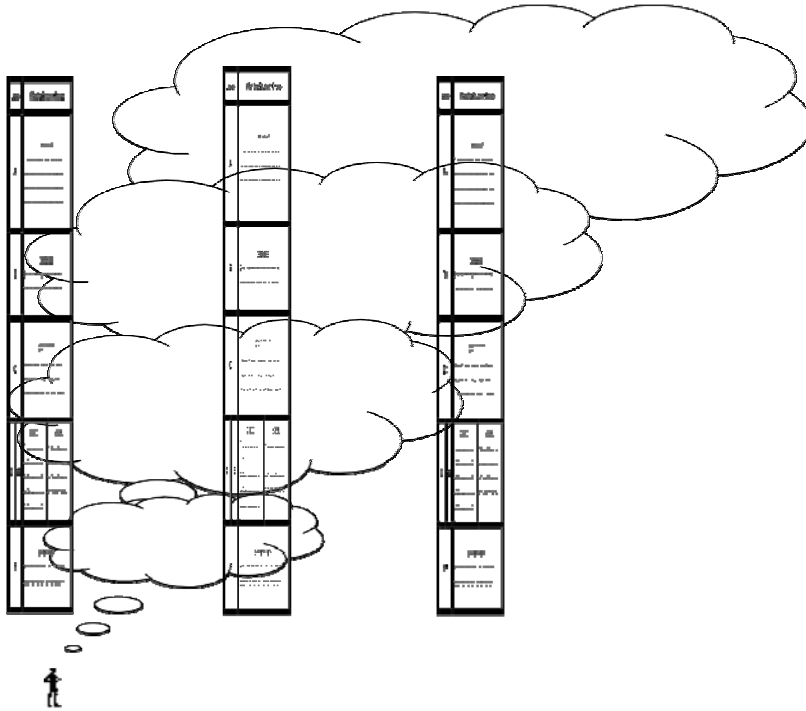


Figure 11. One possible relationship. The levels of the learning progression are levels of several construct maps.

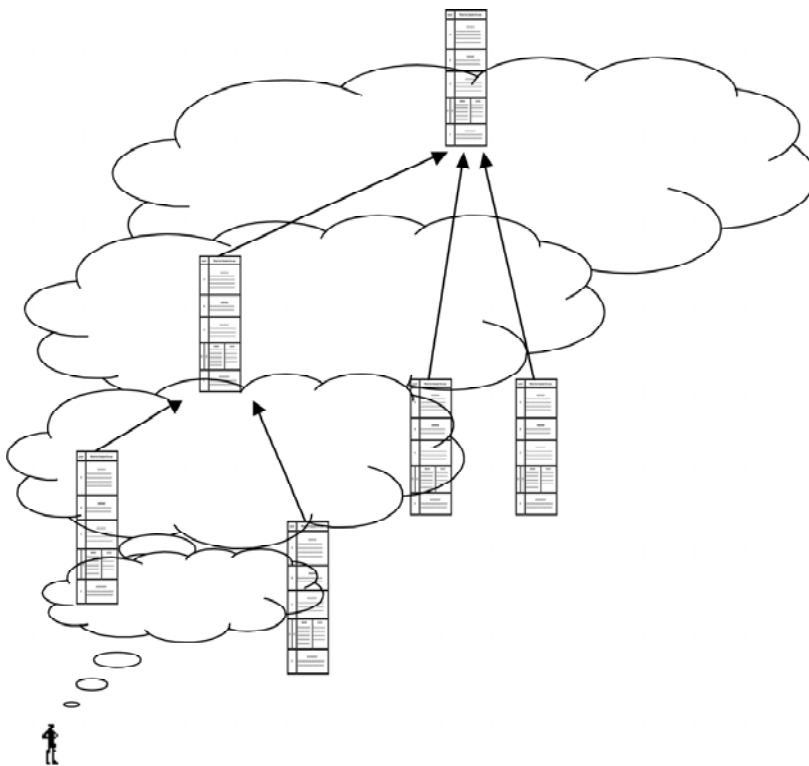


Figure 12. In this situation there is a complicated dependency relationship between the construct maps in the learning progression.

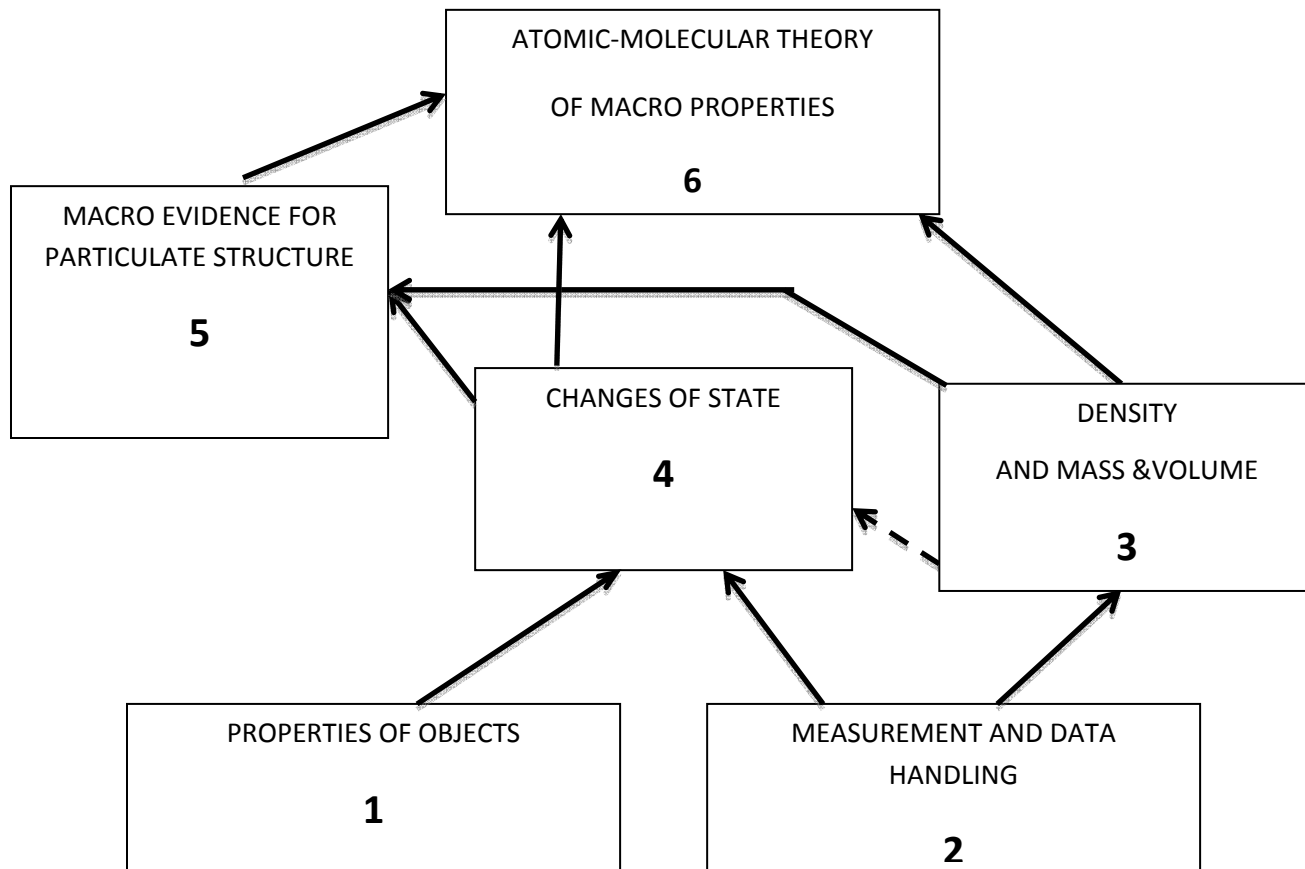


Figure 13. A set of constructs hypothesized to constitute a molecular theory of matter.

Conclusion

In summary, I have argued that the types of accountability systems common at this time have damaged what should be the regular and appropriate connection between the curriculum (as represented by state standards), and instruction and assessment within the classroom, and that this has led to a degradation of their characteristics and functions within the educational system. Specifically, the specification of the state standards that define the state curriculum has led to a particular type of summative assessment that has overwhelmed formative assessment and thus taken over the leading role in guiding classroom instruction. In this way both formative and summative assessments take on roles that in combination narrow the curriculum and hinder good classroom teaching. In contrast, I propose that there is another way that standards can be developed and communicated so that the roles of formative and summative assessment in schooling can be combined (or at the very least rendered complementary and synergistic) and improved, to aid in the success of schools. The key ingredient in creating this synergy is ensuring that (a) there is a common roadmap or learning progression that ties together the standards in an educationally meaningful way, (b) the standards themselves are expressed

and communicated in a richer and more educationally effective way (which are called learning progressions), (c) both forms of assessment (i.e., summative and formative) are based on that learning progression, and (d) the assessments have been built to provide teachers and school administrators with useful information regarding students' placement along the learning progression. However, this alternative cannot consist of just a cosmetic effort to make the standards "look" better. In fact, this program of reform will involve a more determined effort than has been taken on so far. We need to turn the direction of accountability back from its current role, where summative assessment leads formative assessment (and hence to instruction and learning), around to the opposite direction—where instruction and learning, in hand with formative assessment, give the direction to summative assessments and hence to accountability.

References

- American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.
- American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for psychological and educational testing* (2nd ed.). Washington, DC: American Educational Research Association.
- Black, P., & Wilson, M. (2009, April). *Learning progressions to guide systems of formative and summative assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*. New York, NY: Longmans.
- Brown, N. J., Dray, A. J., Lee, Y., & Wilson, M. (2008, March). *Assessing literacy in the classroom: The Berkeley evaluation and assessment system for the striving readers literacy project*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Center for Continuous Instructional Improvement. (2009). *Report of the CCII panel on learning progressions in science* (CPRE Research Report). New York, NY: Columbia University.
- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in Grades K-8*. Washington, DC: National Academies Press.
- German Programme for International Student Assessment Consortium. (2004). *PISA 2003: Ergebnisse des zweiten internationalen Vergleichs* [PISA 2003 Results of the second international comparison]. Münster, Germany: Waxmann.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.

- Kennedy, C. A., Wilson, M., Draney, K., & Tutunciyar, S. (2007). GradeMap 4.2 [Computer software]. Berkeley: University of California Berkeley, Berkeley Evaluation & Assessment Research Center.
- Masters, G. N., Adams, R. A., & Wilson, M. (1990). Charting student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary volume 2* (pp. 628–634). Oxford, England: Pergamon Press.
- Masters, G., & Forster, M. (1996). *Progress maps. Assessment resource kit*. Victoria, Australia: Commonwealth of Australia.
- Organisation for Economic Co-operation and Development. (2005a). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: Author.
- Organisation for Economic Co-operation and Development. (2005b). *PISA 2003 technical report*. Paris, France: Author.
- Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- Perkins, D. (1998). What is understanding? In M.S. Wiske (Ed.), *Teaching for understanding: Linking research with practice*. San Francisco, CA: Jossey-Bass Publishers.
- Reiser, R. A. (2002). A history of instructional design and technology. In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design and technology*. Saddle River, NJ: Prentice-Hall.
- Reiser, B.J., Krajcik, J., Moje, E., & Marx, R. (2003, March). *Design strategies for developing science instructional materials*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.
- Roberts, L., & Sipusic, M. (1999). *Moderation in all things: A class act* [Motion picture]. (Available from the Berkeley Evaluation & Assessment Research Center, Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720-1670)
- Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. New York, NY: Kluwer Academic Publishers.
- Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, 4(1 & 2), 1–98.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433–451.
- Walker, L., Schwartz, R., Dray, A. J., Torres Irribarra, D., Full, M. C., & Wilson, M. (2009, April). *Assessing data modeling*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

- Wilson, M. (2009a, June). *Structured constructs models (SCM): A family of statistical models related to learning progressions*. Paper presented at the learning progressions in science (LeaPS) conference, Iowa City, IA.
- Wilson, M. (2009b). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching*, 46(6), 716–730.
- Wilson, M., & Bertenthal, M. (Eds.). (2006). *Systems for state science assessment*. Washington, DC: National Academies Press.
- Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26(3), 283–306.
- Wilson, M., Scalise, K., Galpern, A., & Lin, Y.-H. (2009). *A guide to the Formative Assessment Delivery System (FADS)*. Berkeley: University of California Berkeley, Berkeley Evaluation & Assessment Research Center.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208.