Assessing Mathematical Proficiency
MSRI Publications
Volume **53**, 2007

Chapter 18
Assessment to Improve Learning in Mathematics: The BEAR Assessment System

MARK WILSON AND CLAUS CARSTENSEN

Introduction

The Berkeley Evaluation and Assessment Research (BEAR) Center has for the last several years been involved in the development of an assessment system, which we call the BEAR Assessment System. The system consists of four principles, each associated with a practical "building block" [Wilson 2005] as well as an activity that helps integrate the four parts together (see the section starting on p. 325). Its original deployment was as a curriculum-embedded system in science [Wilson et al. 2000], but it has clear and logical extensions to other contexts such as in higher education [Wilson and Scalise 2006], in large-scale assessment [Wilson 2005]; and in disciplinary areas, such as chemistry [Claesgens et al. 2002], and the focus of this chapter, mathematics.

In this paper, the four principles of the BEAR Assessment System are discussed, and their application to large-scale assessment is described using an example based on a German assessment of mathematical literacy used in conjunction with the Program for the International Student Assessment [PISA 2005a]; see also Chapter 7, this volume). The BEAR Assessment System is based on a conception of a tight inter-relationship between classroom-level and large-scale assessment [Wilson 2004a; Wilson and Draney 2004]. Hence, in the process of discussing this large-scale application, some arguments and examples will be directed towards classroom-level applications, or, more accurately, towards the common framework that binds the two together [Wilson 2004b].

The Assessment Triangle and the BEAR Approach

Three broad elements on which every assessment should rest are described by the Assessment Triangle from the National Research Council's report *Knowing What Students Know* [NRC 2001] shown in Figure 1.

According to *Knowing What Students Know*, an effective assessment design requires:

- *a model of student cognition and learning* in the field of study;
- well-designed and tested assessment questions and tasks, often called *items*;
- and ways to make *inferences about student competence* for the *particular context of use*. (p. 296)

These elements are of course inextricably linked, and reflect concerns similar to those addressed in the conception of constructive alignment [Biggs 1999], regarding the desirability of achieving goodness-of-fit among learning outcomes, instructional approach, and assessment.

Models of student learning should specify the most important aspects of student achievement to assess, and they provide clues about the types of tasks that will elicit evidence and the types of inferences that can connect observations to learning models and ideas about cognition. To collect responses that serve as high-quality evidence, items themselves need to be systematically developed with both the learning model and the character of subsequent inferences in mind, and they need to be trialed, and the results of the trials systematically examined. Finally, the nature of inferences desired provides the "why" of it all — if we don't know what we want to do with the assessment information, then we can't figure out what the student model or the items should be. Of course, context determines many specifics of the assessment.

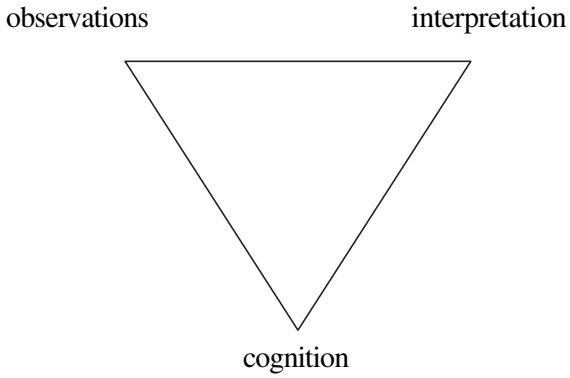


Figure 1. The *Knowing What Students Know* assessment triangle.

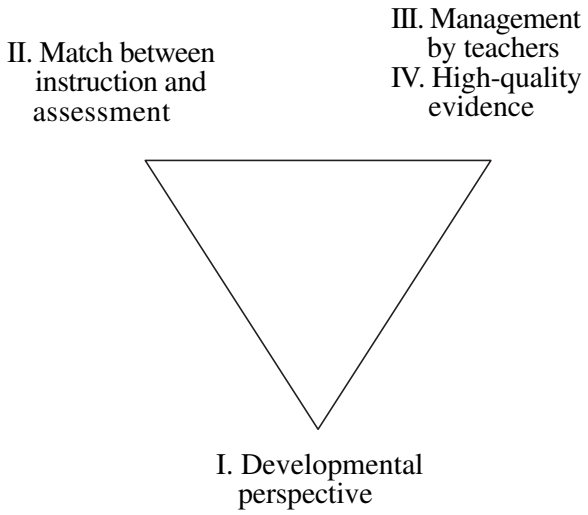


Figure 2. The principles of the BEAR assessment system.

The BEAR Assessment System is based on the idea that good assessment addresses these considerations through four principles: (1) a developmental perspective, (2) a match between instruction and assessment, (3) the generating of high-quality evidence, and (4) management by instructors to allow appropriate feedback, feed-forward, and follow-up. Connections between these principles and the assessment triangle are illustrated in Figure 2. See [Wilson 2005] for a detailed account of an instrument development process based on these principles. Next, we discuss each of these principles and their implementation.

Principle 1: Developmental Perspective

A "developmental perspective" regarding student learning means assessing the development of student understanding of particular concepts and skills over time, as opposed to, for instance, making a single measurement at some final or supposedly significant time point. Criteria for developmental perspectives have been challenging goals for educators for many years. What to assess and how to assess it, whether to focus on generalized learning goals or domain-specific knowledge, and the implications of a variety of teaching and learning theories all impact what approaches might best inform developmental assessment. From Bruner's nine tenets of hermeneutic learning [Bruner 1996] to considerations of empirical, constructivist, and sociocultural schools of thought [Olson and Torrance 1996] to the recent National Research Council report *How People Learn* [NRC 2000], broad sweeps of what might be considered in a developmental

perspective have been posited and discussed. Cognitive taxonomies such as Bloom's Taxonomy of Educational Objectives [1956], Haladyna's Cognitive Operations Dimensions [1994] and the Structure of the Observed Learning Outcome (SOLO) Taxonomy [Biggs and Collis 1982] are among many attempts to concretely identify generalizable frameworks. One issue is that as learning situations vary, and their goals and philosophical underpinnings take different forms, a "one-size-fits-all" development assessment approach rarely satisfies course needs. Much of the strength of the BEAR Assessment System comes in providing tools to model many different kinds of learning theories and learning domains. What is to be measured and how it is to be valued in each BEAR assessment application is drawn from the expertise and learning theories of the teachers and/or curriculum developers involved in the developmental process.

Building block 1: Progress variables. Progress variables [Masters et al. 1990; Wilson 1990] embody the first of the four principles: that of a developmental perspective on assessment of student achievement and growth. The four building blocks and their relationship to the assessment triangle are shown in Figure 3. The term "variable" is derived from the measurement concept of focusing on one characteristic to be measured at a time. A progress variable is a well-thought-out and researched ordering of qualitatively different levels of performance. Thus, a variable defines what is to be measured or assessed in terms general enough to be interpretable at different points in a curriculum but specific enough to guide the development of the other curriculum and assessment components. When the goals of the instruction are linked to the set of variables, then the set of variables also define what is to be taught. Progress variables are one model of

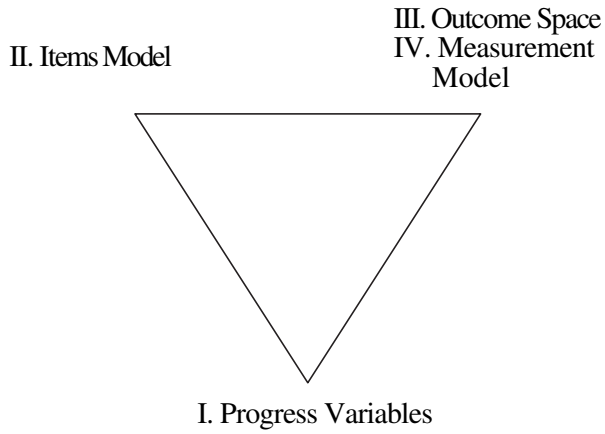


Figure 3. The building blocks of the BEAR assessment system.

how assessments can be connected with instruction and accountability. Progress variables provide a way for large-scale assessments to be linked in a principled way to what students are learning in classrooms, while remaining independent of the content of a specific curriculum.

The approach assumes that, within a given curriculum, student performance on progress variables can be traced over the course of the year, facilitating a more developmental perspective on student learning. Assessing the growth of students' understanding of particular concepts and skills requires a model of how student learning develops over a set period of (instructional) time. A growth perspective helps one to move away from "one-shot" testing situations, and away from cross-sectional approaches to defining student performance, toward an approach that focuses on the process of learning and on an individual's progress through that process. Clear definitions of what students are expected to learn, and a theoretical framework of how that learning is expected to unfold, as the student progresses through the instructional material, are necessary to establish the construct validity of an assessment system.

Explicitly aligning the instruction and assessment addresses the issue of the content validity[1] of the assessment system as well. Traditional testing practices — in standardized tests as well as in teacher-made tests — have long been criticized for oversampling items that assess only basic levels of knowledge of content and ignore more complex levels of understanding. Relying on progress variables to determine what skills are to be assessed means that assessments focus on what is important, not what is easy to assess. Again, this reinforces the central instructional objectives of a course. Resnick and Resnick [1992, p. 59] have argued: "Assessments must be designed so that when teachers do the natural thing — that is, prepare their students to perform well — they will exercise the kinds of abilities and develop the kinds of skill and knowledge that are the real goals of educational reform." Variables that embody the aims of instruction (e.g., "standards") can guide assessment to do just what the Resnicks were demanding. In a large-scale assessment, the notion of a progress variable will be more useful to the parties involved than simple number-correct scores or standings relative to some norming population, providing the diagnostic information so often requested (see also Chapters 10, 12, 14, 21 and 22 of this volume.)

The idea of using variables (note that, for the sake of brevity, I will refer to these as "variables") also offers the possibility of gaining significant *efficiency* in assessment: Although each new curriculum prides itself on bringing something new to the subject matter, in truth, most curricula are composed of a common

[1] Content validity is evidence that the content of an assessment is a good representation of the construct it is intended to cover (see [Wilson 2005, Chapter 8]).

stock of content. And, as the influence of national and state standards increases, this will become more true, and also easier to codify. Thus, we might expect innovative curricula to have one, or perhaps even two progress variables that do not overlap with typical curricula, but the remainder will form a fairly stable set that will be common across many curricula.

Progress variables are derived in part from research into the underlying cognitive structure of the domain and in part from professional opinion about what constitutes higher and lower levels of performance or competence, but are also informed by empirical research into how students respond to instruction or perform in practice [NRC 2001]. To more clearly understand what a progress variable is, let us consider an example.

The example explored in this chapter is a test of mathematics competency taken from one booklet of a German mathematics test administered to a random subsample of the German PISA sample of 15-year-old students in the 2003 administration [PISA 2004]. The test was developed under the same general guidelines as the PISA mathematics test (see Chapter 7 in this volume), where Mathematical Literacy is a "described variable" (i.e., the PISA jargon for progress variable) with several successive levels of sophistication in performing mathematical tasks [PISA 2005a; 2005b]. These levels are as follows:

PISA Levels of Mathematical Literacy

VI. At Level VI students can conceptualize, generalize, and utilize information based on their investigations and modeling of complex problem situations. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply their insight and understandings along with a mastery of symbolic and formal mathematical operations and relationships to develop new approaches and strategies for attacking novel situations. Students at this level can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situations.

V. At Level V students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterizations, and insight pertaining to these situations. They can reflect on their actions and formulate and communicate their interpretations and reasoning.

IV. At Level IV students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilize well-developed skills and reason flexibly, with some insight, in these contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments, and actions.

III. At Level III students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.

II. At Level II students can interpret and recognize situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions. They are capable of direct reasoning and making literal interpretations of the results.

I. At Level I students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli.

The levels shown above were derived from a multistep process [PISA 2005b] as follows: (a) Mathematics curriculum experts identified possible subscales in the domain of mathematics, (b) PISA items were mapped onto each subscale, (c) a skills audit of each item in each subscale was carried out on the basis of a detailed expert analysis, (d) field test data were analyzed to yield item locations on subscales, (e) 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 possible values of the progress variable. This results in natural clusters of skills, which provides a basis for understanding and describing the progress variable. The results of this last step were also validated with later empirical data, and by using a validation process involving experts. Note that this method of developing a progress variable is much less precise than the approaches described in the references above (e.g., [Wilson et al. 2000; Wilson and Scalise 2006], and will thus usually result in a progress variable that is much broader in its content.

Principle 2: Match Between Instruction and Assessment

The match between instruction and assessment in the BEAR Assessment System is established and maintained through two major parts of the system: progress variables, described above, and assessment tasks or activities, described in this section. The main motivation for the progress variables so far developed is that they serve as a framework for the assessments and a method of making measurement possible. However, this second principle makes clear that the framework for the assessments and the framework for the curriculum and instruction must be one and the same. This is not to imply that the needs of assessment must drive the curriculum, nor that the curriculum description will entirely determine the assessment, but rather that the two, assessment and instruction, must be in step — they must both be designed to accomplish the same thing, the aims of learning, whatever those aims are determined to be.

Using progress variables to structure both instruction and assessment is one way to make sure that the two are in alignment, at least at the planning level. In order to make this alignment concrete, however, the match must also exist at the level of classroom interaction and that is where the nature of the assessment tasks becomes crucial. Assessment tasks need to reflect the range and styles of the instructional practices in the curriculum. They must have a place in the "rhythm" of the instruction, occurring at places where it makes instructional sense to include them, usually where instructors need to see how much progress their students have made on a specific topic. See [Minstrell 1998] for an insightful account of such occasions.

One good way to achieve this is to develop both the instructional materials and the assessment tasks at the same time — adapting good instructional sequences to produce assessable responses and developing assessments into full-blown instructional activities. Doing so brings the richness and vibrancy of curriculum development into assessment, and also brings the discipline and hard-headedness of explaining assessment data into the design of instruction.

By developing assessment tasks as part of curriculum materials, they can be made directly relevant to instruction. Assessment can become indistinguishable from other instructional activities, without precluding the generation of high-quality, comparative, and defensible assessment data on individual students and classes.

Building block 2: The items design. The items design governs the match between classroom instruction and the various types of assessment. The critical element to ensure this in the BEAR assessment system is that each assessment task is matched to at least one variable.

A variety of different task types may be used in an assessment system, based

on the requirements of the particular situation. There has always been a tension in assessment situations between the use of multiple-choice items, which are perceived to contribute to more reliable assessment, and other, alternative forms of assessment, which are perceived to contribute to the validity of a testing situation. The BEAR Assessment System includes designs that use different item types to resolve this tension.

When using this assessment system within a curriculum, a particularly effective mode of assessment is what we call *embedded assessment*. By this we mean that opportunities to assess student progress and performance are integrated into the instructional materials and are virtually indistinguishable from the day-to-day classroom activities. We found it useful to think of the metaphor of a stream of instructional activity and student learning, with the teacher dipping into the stream of learning from time to time to evaluate student progress and performance. In this model or metaphor, assessment becomes *part* of the teaching and learning process, and we can think of it being assessment for learning [Black et al. 2003]. If assessment is also a learning event, then it does not take unnecessary time away from instruction, *and* the number of assessment tasks can be more efficiently increased in order to improve the reliability of the results [Linn and Baker 1996]. But, for assessment to become fully and meaningfully embedded in the teaching and learning process, the assessment must be linked to a specific curriculum, i.e. it must be curriculum dependent, not curriculum independent as must be the case in many high-stakes testing situations [Wolf and Reardon 1996].

In embedded assessment in classrooms, there will be a variety of different types of assessment tasks, just as there is variety in the instructional tasks. These may include individual and group "challenges," data processing questions, questions following student readings, and even instruction/assessment events such as "town meetings." Such tasks may be constructed-response, requiring students to fully explain their responses in order to achieve a high score, or they may be multiple choice, freeing teachers from having to laboriously hand score all of the student work [Briggs et al. 2006].

There are many variations in the way that progress variables can be made concrete in practice, from using different assessment modes (multiple choice, performance assessment, mixed modes, etc.), to variations in the frequency of assessing students (once a week, once a month, etc.), to variations in the use of embedding of assessments (all assessments embedded, some assessments in a more traditional testing format, etc.).

In large-scale testing situations, the basis on which the mix of assessment modes is decided may be somewhat different from that in embedded assessment contexts. Many large-scale tests are subject to tight constraints both in terms of

the time available for testing, and in terms of the financial resources available for scoring. Thus, although performance assessments are valued because of their perceived high validity, it may not be possible to collect enough information through performance assessments alone to accurately estimate each examinee's proficiency level; multiple-choice items, which require less time to answer and which may be scored by machine rather than by human raters, may be used to increase the reliability of the large-scale test.

Returning to the German Mathematical Literacy example, the test booklet contained 64 dichotomous items; 18 of these items were selected for this example. Examples of these items are the tasks Function, Rectangle and Difference, shown on the next page. 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 Function. Numerical Modeling requires the students to construct solutions for problems with given numbers in one or more steps — see the item Rectangle. 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 Difference. Because the collection of items follows an experimental design, 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 selected set of items has a 3×3 design with two observations of each pair of conditions, resulting in 18 items in total.

Principle 3: Management by Teachers

For information from the assessment tasks and the BEAR analysis to be useful to instructors and students, it must be couched in terms that are directly related to the instructional goals associated with the progress variables. Constructed response tasks, if used, must be quickly, readily, and reliably scorable. The categories into which the scores are sorted must be readily interpreted in an educational setting, whether it is within a classroom, by a parent, or in a policy-analysis setting. The requirement for transparency in the relationship between scores and actual student responses to an item leads to the third building block.

Example tasks from the German Mathematical Literacy booklet

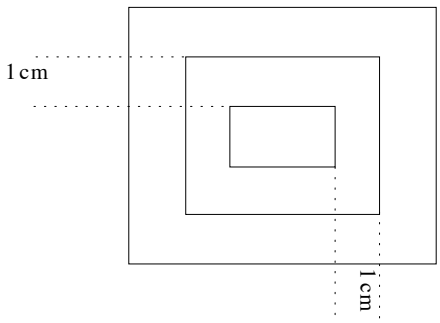
Function (a technical processing item in algebra)

Consider the function given by the equation $y = 2x - 1$. Fill in the missing values.

x	-2	-1	0		3	$\cdots$		
y						$\cdots$	19	

Rectangles (a numerical modeling item in algebra)

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



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

The length increases by ___ cm. The width increases by ___ cm.

The perimeter increases by ___ cm.

Difference (an abstract modeling item in algebra)

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.)

first number ☐☐☐

second number ☐☐☐

Building block 3: The outcome space. The outcome space is the set of outcomes into which student performances are categorized for all the items associated with a particular progress variable. In practice, these are presented as scoring guides for student responses to assessment tasks. This is the primary means by which the essential element of teacher professional judgment is implemented in the BEAR Assessment System. These are supplemented by "exemplars": examples of student work at every scoring level for every task and variable combination, and "blueprints," which provide the teachers with a layout showing opportune times in the curriculum to assess the students on the different progress variables.

For the information from assessment opportunities to be useful to teachers, it must be couched in terms that are directly interpretable with respect to the instructional goals associated with the progress variables. Moreover, this must be done in a way that is intellectually and practically efficient. Scoring guides have been designed to meet these two criteria. A scoring guide serves as a operational definition for a progress variable by describing the performance criteria necessary to achieve each score level of the variable.

The scoring guides are meant to help make the performance criteria for the assessments clear and explicit (or "transparent and open" to use Glaser's [1990] terms) — not only to the teachers but also to the students and parents, administrators, or other "consumers" of assessment results. In fact, we strongly recommend to teachers that they share the scoring guides with administrators, parents and students, as a way of helping them understand what types of cognitive performance were expected and to model the desired processes.

In addition, students appreciate the use of scoring guides in the classroom. In a series of interviews with students in a Kentucky middle school that was using the BEAR Assessment System (reported in [Roberts and Sipusic 1999]), the students spontaneously expressed to us their feeling that, sometimes for the first time, they understood what it was that their teachers expected of them, and felt they knew what was required to get a high score. The teachers of these students found that the students were often willing to redo their work in order to merit a higher score.

Traditional multiple-choice items are, of course, based on an implicit scoring guide — one option is correct, the others all incorrect. Alternative types of multiple-choice items can be constructed that are explicitly based on the levels of a construct map [Briggs et al. 2006], and thus allow a stronger interpretation of the test results. For the German Mathematical Literacy example, the items are all traditional multiple choice — their development did not involve the explicit construction of an outcome space.

Principle 4: High-Quality Evidence

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure values of a progress variable as described above, or even to develop a reasonable framework that can be supported by evidence. To ensure comparability of results across time and context, procedures are needed to (a) examine the coherence of information gathered using different formats, (b) map student performances onto the progress variables, (c) describe the structural elements of the accountability system — tasks and raters — in terms of the progress variables, and (d) establish uniform levels of system functioning, in terms of quality control indices such as reliability. Although this type of discussion can become very technical to consider, it is sufficient to keep in mind that the traditional elements of assessment standardization, such as validity/reliability studies and bias/equity studies, must be carried out to satisfy quality control and ensure that evidence can be relied upon.

Building block 4: Wright maps. Wright maps represent the principle of high-quality evidence. Progress maps are graphical and empirical representations of a progress variable, showing how it unfolds or evolves in terms of increasingly sophisticated student performances. They are derived from empirical analyses of student data on sets of assessment tasks. Maps are based on an ordering of these assessment tasks from relatively easy tasks to more difficult and complex ones. A key feature of these maps is that both students and tasks can be located on the same scale, giving student proficiency the possibility of substantive interpretation, in terms of what the student knows and can do and where the student is having difficulty. The maps can be used to interpret the progress of one particular student, or the pattern of achievement of groups of students, ranging from classes to nations.

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 [PISA 2005a]. A Wright map illustrating the estimates for the Rasch model is shown in Figure 4. On this map, an "X" represents a group of students, all at the same estimated achievement level. 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" each represent a Technical Processing, Numerical Modeling, and Abstract Modeling item, with the Topic Area indicated by the column headings above. Where a student is located 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

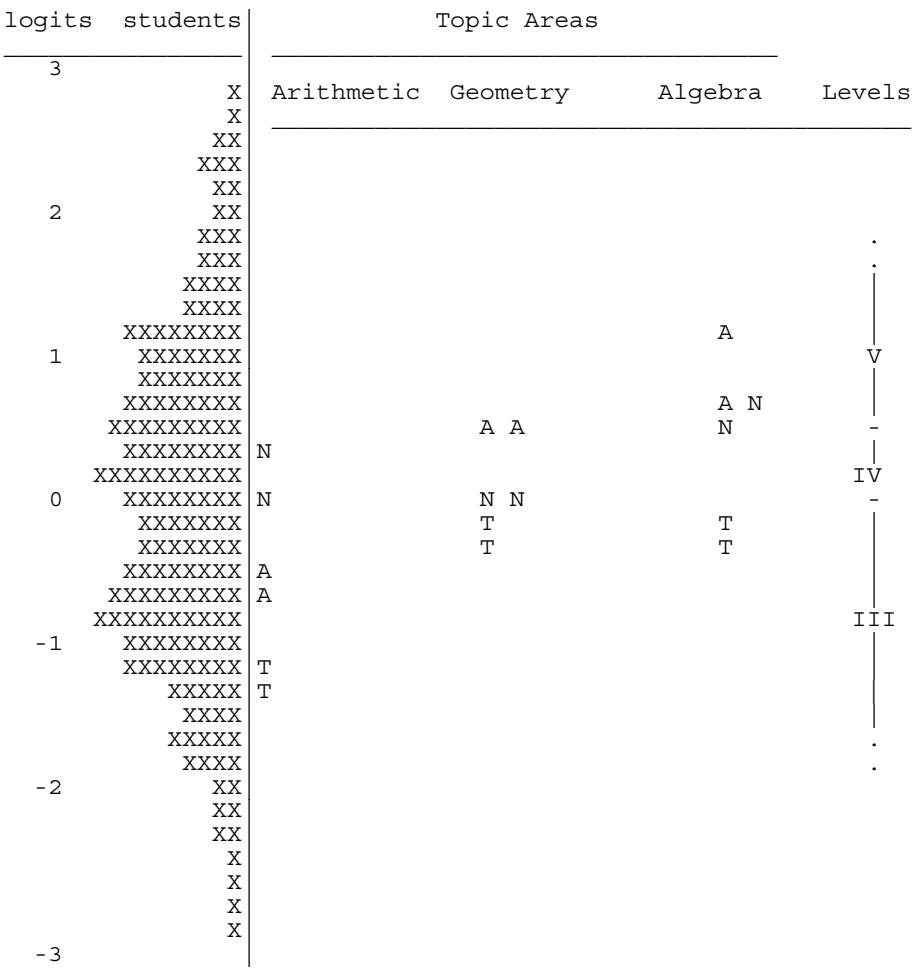


Figure 4. A Wright map of the mathematical literacy variable.

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 this map illustrates the description of the Mathematical Literacy progress variable in terms of the Levels from page 316 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 topic that deserves a follow-up investigation.

We typically use a multi-dimensional Rasch modeling approach to calibrate the maps for use in the BEAR Assessment System (see [Adams et al. 1997] for the specifics of this model). These maps have at least two advantages over the traditional method of reporting student performance as total scores or percentages: First, it allows teachers to interpret a student's proficiency in terms of average or typical performance on representative assessment activities; and second, it takes into consideration the relative difficulties of the tasks involved in assessing student proficiency.

Once constructed, maps can be used to record and track student progress and to illustrate the skills that students have mastered and those that the students are working on. By placing students' performance on the continuum defined by the map, teachers, administrators, and the public can interpret student progress with respect to the standards that are inherent in the progress variables. Wright maps can come in many forms, and have many uses in classroom and other educational contexts. In order to make the maps flexible and convenient enough for use by teachers and administrators, we have also developed software for teachers to use to generate the maps. This software, which we call *GradeMap* [Kennedy et al. 2005], allows consumers to enter the scores given to the students on assessments, and then map the performance of groups of students, either at a particular time or over a period of time.

Bringing It All Together: Performance Standard Setting

The final ingredient in the BEAR Assessment System is the means by which the four building blocks discussed thus far are brought together into a coherent system — in the case of large-scale assessment, by standard setting. We have developed a standard-setting procedure, called "construct mapping" [Wilson and Draney 2002], that allows the standard-setting committee members to use the item response map as a model of what a student at a given level knows and can do. The map is represented in a piece of software [Hoskens and Wilson 1999] that allows standard-setting committee members to find out about the details of student performance at any given proficiency level, and to assist them in deciding where the cutoffs between performance levels should be.

An example showing a section of such an item map is given in Figure 5. This illustration is from a somewhat more complicated example than the German Mathematical Literacy Test, involving both multiple-choice items and two items that required written responses (WR1 and WR2) and which were scored into five ordered categories [Wilson 2005]. The column on the far left contains a numerical scale that allows the selection and examination of a given point on the map, and the selection of the eventual cut scores for the performance levels. This scale is a transformation of the original logit scale, designed to

	Multiple Choice		WR 1		WR 2	
Scale		P		P		P
620						
610						
600						
590						
580						
570	37	.30			2.3	.26
560	15	.34				
550						
540	28 39	.38				
530	27	.41				
520	19 38	.45				
510						
500	34 43 45 48	.50	1.3	.40		
490	17 18 20 40 50	.53				
480	4 31	.56				
470	11 32 33 44 47	.59				
460	5 9 12 46	.61				
450	3 6 7 10 16 29	.64				
440	36	.67				
430	8 14 22 23 26 35	.69				
420	13 24 25	.71				
410	41 42	.73				
400	1 21 30 49	.76				
390						
380					2.2	.56
370	2	.82				
360			1.2	.40		
350						

Figure 5. A screen-shot of a cut-point setting map.

have a mean of 500, and to range from approximately 0 to 1000. (This choice of scale is a somewhat arbitrary, but designed to avoid negative numbers and small decimals, which some members of standard-setting committees find annoying.) The next two columns contain the location of the multiple-choice items (labeled by number of appearance on the examination), and the probability that a person at the selected point would get each item correct (in this case, a person at 500 on the scale — represented by the shaded band across the map). The next two sets of columns display the thresholds for the two written-response items — for example, the threshold levels for scores of 2 and 3 on written-response item 1 are represented by 1.2 and 1.3, respectively (although each item is scored on a scale of 1 to 5 on this particular examination, only the part of the scale where a person would be most likely to get a score of 2 or 3 on either item is shown) — and the probability that a person at 500 on the scale would score at that particular score level on each item. The software also displays, for a person at the selected point on the logit scale, the expected score total on the multiple-choice section (Figure 5 does not show this part of the display), and the expected score on each of the written response items.

In order to set the cut points, the committee first acquaints itself with the test materials. The meaning of the various parts of the map is then explained, and the committee members and the operators of the program spend time with the software familiarizing themselves with points on the scale.

The display of multiple-choice item locations in ascending difficulty, next to the written-response thresholds, helps to characterize the scale in terms of what increasing proficiency “looks like” in the pool of test-takers. For example, if a committee were considering 500 as a cut point between performance levels, it could note that 500 is a point at which items like 34, 43, 45, and 48 are expected to be chosen correctly about 50% of the time, a harder item like 37 is expected to be chosen correctly about 30%, and easier items like 2 are expected to be chosen correctly 80% of the time. The set of multiple-choice items, sorted so they are in order of ascending difficulty, is available to the committee so that the members can relate these probabilities to their understanding of the items. The committee could also note that a student at that point (i.e., 500), would be equally likely to score a 2 or a 3 on the first written-response item (40% each) and more likely to score a 2 than a 3 on the second (56% vs. 26%). Examples of student work at these levels would be available to the committee for consideration of the interpretation of these scores. Committee members can examine the responses of selected examinees to both the multiple-choice and written-response items, chart their locations on the map, and judge their levels.

The committee then, through a consensus-building process, sets up cut points on this map, using the item response calibrations to allow interpretation in

terms of predicted responses to both multiple-choice items and open-ended constructed-response items. Locations of an individual student's scores and distributions of the scaled values of the progress variable are also available for interpretative purposes. This procedure allows both criterion-referenced and norm-referenced interpretations of cut scores.

Use of the maps available from the item response modeling approach not only allows the committees to interpret cut-offs in a criterion-referenced way, it also allows maintenance of similar standards from year to year by equating of the item response scales. This can be readily accomplished by using linking items on successive tests to keep the waves of data on the same scale — hence the cut-offs set one year can be maintained in following years.

Discussion

A central tenet of the assessment reforms of recent years ("authentic," "performance," etc.) has been the WYTIWYG principle — "What you test is what you get." This principle has led the way for assessment reform at the state or district level nationwide. The assumption behind this principle is that assessment reforms will not only affect assessments *per se*, but these effects will trickle down into the curriculum and instruction that students receive in their daily work in classrooms. Hence, when one looks to the curricula that students are experiencing, one would expect to see such effects, and, in particular, one would expect to see these effects even more strongly in the cutting-edge curricula that central research agencies such as the U.S. National Science Foundation (NSF) sponsor. Thus it is troubling to find that this does not seem to be the case: An NSF review of new middle school science curricula [NSF 1997] found only one where the assessment itself reflected the recent developments in assessment. For that one (the *IEY Assessment System* — see [Wilson et al. 2000]), it was found that the reformed assessment did indeed seem to have the desired sorts of effects [Wilson and Sloane 2000], but for the other curricula no such effects were possible, because the assessment reforms have not, in general, made it into them.

We have demonstrated a way in which large-scale assessments can be more carefully linked to what students are learning. The key here is the use of progress variables to provide a common conceptual framework across curricula. Variables developed and used in the ways we have described here can mediate between the level of detail that is present in the content of specific curricula and the necessarily more vague contents of standards documents. This idea of a "crosswalk between standards and assessments" has also been suggested by Eva Baker of the Center for Research on Evaluation, Standards, and Student Testing [Land 1997, p. 6]. These variables also create a "conceptual basis" for relating

a curriculum to standards documents, to other curricula, and to assessments that are not specifically related to that curriculum.

With the assessments to be used across curricula structured by progress variables, the problem of item development is lessened — ideas and contexts for assessment tasks may be adapted to serve multiple curricula that share progress variables. The cumulative nature of the curricula is expressed through (a) the increasing difficulty of assessments and (b) the increasing sophistication needed to gain higher scores using the assessment scoring guides. Having the same underlying structure makes clear to teachers, policy-makers, and parents what is the ultimate purpose of each instructional activity and each assessment, and also makes easier the diagnostic interpretation of student responses to the assessments.

The idea of a progress variable is not radically new — it has grown out of the traditional approach to test content — most tests have a "blueprint" or plan that assigns items to particular categories, and hence, justifies why certain items are there, and others aren't. The concept of a progress variable goes beyond this by looking more deeply into why we use certain assessments when we do (i.e., by linking them to growth through the curriculum), and by calibrating the assessments with empirical information.

Although the ideas inherent in components of the BEAR Assessment System are not unique, the combination of these particular ideas and techniques into a usable system does represent a new step in assessment development. The implications for this effort for other large-scale tests, for curricula, and for assessment reform on a broader level, need to be explored and tested through other related efforts. We hope our efforts and experiences will encourage increased discussion and experimentation of the use of state of the art assessment procedures across a broad range of contexts from classroom practice to large-scale assessments.

References

[Adams et al. 1997] R. J. Adams, M. Wilson, and W.-C. Wang, "The multidimensional random coefficients multinomial logit model", *Applied Psychological Measurement* **21**:1 (1997), 1–23.

[Biggs 1999] J. B. Biggs, *Teaching for quality learning at university*, Buckingham: SRHE and Open University Press, 1999.

[Biggs and Collis 1982] J. B. Biggs and K. F. Collis, *Evaluating the quality of learning: The SOLO taxonomy*, New York: Academic Press, 1982.

[Black et al. 2003] P. Black, C. Harrison, C. Lee, B. Marshall, and D. Wiliam, *Assessment for learning*, London: Open University Press, 2003.

[Bloom 1956] B. S. Bloom (editor), *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*, New York and Toronto: Longmans, Green, 1956.

[Briggs et al. 2006] D. Briggs, A. Alonzo, C. Schwab, and M. Wilson, "Diagnostic assessment with ordered multiple-choice items", *Educational Assessment* **11**:1 (2006), 33–63.

[Bruner 1996] J. Bruner, *The culture of education*, Cambridge, MA: Harvard University Press, 1996.

[Claesgens et al. 2002] J. Claesgens, K. Scalise, K. Draney, M. Wilson, and A. Stacy, "Perspectives of chemists: A framework to promote conceptual understanding of chemistry", paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 2002.

[Glaser 1990] R. Glaser, *Testing and assessment: O tempora! O mores!*, Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1990.

[Haladyna 1994] T. M. Haladyna, "Cognitive taxonomies", pp. 104–110 in *Developing and validating multiple-choice test items*, edited by T. M. Haladyna, Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.

[Hoskens and Wilson 1999] M. Hoskens and M. Wilson, *StandardMap* [Computer program], Berkeley, CA: University of California, 1999.

[Kennedy et al. 2005] C. A. Kennedy, M. Wilson, and K. Draney, *GradeMap 4.1* [Computer program], Berkeley, California: Berkeley Evaluation and Assessment Center, University of California, 2005.

[Land 1997] R. Land, "Moving up to complex assessment systems", *Evaluation Comment* **7**:1 (1997), 1–21.

[Linn and Baker 1996] R. Linn and E. Baker, "Can performance-based student assessments be psychometrically sound?", pp. 84–103 in *Performance-based student assessment: Challenges and possibilities. Ninety-fifth Yearbook of the National Society for the Study of Education*, edited by J. B. Baron and D. P. Wolf, Chicago: University of Chicago Press, 1996.

[Masters et al. 1990] G. N. Masters, R. A. Adams, and M. Wilson, "Charting student progress", pp. 628–634 in *International encyclopedia of education: Research and studies*, vol. 2 (Supplementary), edited by T. Husen and T. N. Postlethwaite, Oxford and New York: Pergamon, 1990.

[Minstrell 1998] J. Minstrell, "Student thinking and related instruction: Creating a facet-based learning environment", paper presented at the meeting of the Committee on Foundations of Assessment, Woods Hole, MA, October 1998.

[NRC 2000] National Research Council (Committee on Developments in the Science of Learning, Commission on Behavioral and Social Sciences and Education), *How people learn: Brain, mind, experience, and school*, expanded ed., edited by J. D. Bransford et al., Washington, DC: National Academy Press, 2000.

[NRC 2001] National Research Council Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education), *Knowing what students know: The science and design of educational assessment*, edited by J. Pelligrino et al., Washington, DC: National Academy Press, 2001.

[NSF 1997] National Science Foundation, *Review of instructional materials for middle school science*, Arlington, VA: Author, 1997.

[Olson and Torrance 1996] D. R. Olson and N. Torrance (editors), *Handbook of education and human development: New models of learning, teaching and schooling*, Oxford: Blackwell, 1996.

[PISA 2004] PISA, *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland: Ergebnisse des zweiten internationalen Vergleichs*, Münster: Waxmann, 2004.

[PISA 2005a] Programme for International Student Assessment, "Learning for tomorrow's world: First results from PISA 2003", Technical report, Paris: Organisation for Economic Co-operation and Development, 2005.

[PISA 2005b] Programme for International Student Assessment, "PISA 2003 Technical Report", Technical report, Paris: Organisation for Economic Co-operation and Development, 2005.

[Resnick and Resnick 1992] L. B. Resnick and D. P. Resnick, "Assessing the thinking curriculum: New tools for educational reform", pp. 37–76 in *Changing assessments*, edited by B. R. Gifford and M. C. O'Connor, Boston: Kluwer, 1992.

[Roberts and Sipusic 1999] L. Roberts (producer) and M. Sipusic (director), "Moderation in all things: A class act" [Film], available from the Berkeley Evaluation and Assessment Center, Graduate School of Education, University of California, Berkeley, CA 94720–1670, 1999.

[Wilson 1990] M. Wilson, "Measurement of developmental levels", pp. 628–634 in *International encyclopedia of education: Research and studies*, vol. Supplementary vol. 2, edited by T. Husen and T. N. Postlethwaite, Oxford: Pergamon Press, 1990.

[Wilson 2004a] M. Wilson, "A perspective on current trends in assessment and accountability: Degrees of coherence", pp. 272–283 in [Wilson 2004b], 2004.

[Wilson 2004b] M. Wilson (editor), *Towards coherence between classroom assessment and accountability: One hundred and third yearbook of the National Society for the Study of Education*, part 2, Chicago: University of Chicago Press, 2004.

[Wilson 2005] M. Wilson, *Constructing measures: An item response modeling approach*, Mahwah, NJ: Lawrence Erlbaum Associates, 2005.

[Wilson and Draney 2002] M. Wilson and K. Draney, "A technique for setting standards and maintaining them over time", pp. 325–332 in *Measurement and multivariate analysis*, edited by S. Nishisato et al., Tokyo: Springer-Verlag, 2002.

[Wilson and Draney 2004] M. Wilson and K. Draney, "Some links between large-scale and classroom assessments: The case of the BEAR Assessment System", pp. 132–154 in [Wilson 2004b], 2004.

[Wilson and Scalise 2006] M. Wilson and K. Scalise, "Assessment to improve learning in higher education: The BEAR Assessment System", *Higher Education* **52**:4 (2006), 635–663.

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

[Wilson et al. 2000] M. Wilson, L. Roberts, K. Draney, and K. Sloane, *SEPUP assessment resources handbook*, Berkeley, CA: Berkeley Evaluation and Assessment Research Center, University of California, 2000.

[Wolf and Reardon 1996] D. P. Wolf and S. Reardon, "Access to excellence through new forms of student assessment", pp. 52–83 in *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education*, edited by J. B. Baron and D. P. Wolf, Chicago: University of Chicago Press, 1996.