On coherence and core ideas

Mark Wilson & Karen Draney

University of California, Berkeley

August 2009

Paper prepared for the Organizing K-12 Science Education around Core Ideas meeting organized by the National Research Council (NRC) Board on Science Education , August 17, 2009.

Introduction

The aim of this paper is to summarize some perspectives on what is needed in order to develop and agree to core ideas in science education, in particular, as these relate to the possibilities of constructing and effectively using assessments of those core areas. These perspectives have been developing over a period of approximately twenty years, starting with initial work on the BEAR Assessment System (BAS, Wilson & Sloane, 2000), and then through involvement in some National Research Council committees (NRC, 2001; NRC 2006), as well as a series of assessment development projects funded by the National Science Foundation, the Institute for Education Research and the California State Department of Education.

The central concept on which all of the remarks are based is that assessments can only play a useful and effective role where they are *coherent* with the educational structures in which they must function. Where they are not in such a coherent relationship, their effects, at both the system level and the individual level, may be at best null, or, at times, quite strongly negative. Thus, the development and organization of core ideas must equally engage in these coherences, and the successes and failures of the core ideas will hinge crucially on the understanding that we have of these coherences. The paper will outline the various modes of coherence that we see as being, and exemplify the respective coherences, and at the end will discuss overall perspectives and comment on relations to the policy world.

Different modes of coherence

A successful system of standards-based education needs to be coherent, and this coherence is composed of a number of aspects. First, it needs to be *vertically coherent* in the sense that there is a shared understanding at all levels of the system (classroom, school, school district, and state) of the goals for education that underlie the standards, as well as consensus about the purposes and uses of assessment. Second, the system also needs to be *developmentally coherent*, in the sense that it takes into account what is known about how students' understanding develops over time and the content knowledge, abilities, and understanding that are needed for learning to progress at each stage of the process. Third, it needs to be *horizontally coherent* in the sense that curriculum, instruction, and assessment are all aligned with the standards, target the same goals for learning, and work together to support students' developing science literacy. The definitions of these three types of coherence are shown in Figure 1.

=================
Insert Figure 1 about here
=================

Coherence is necessary in the interrelationship of all the elements of the system. For example, the preparation of beginning teachers and the ongoing professional development of experienced ones should be guided by the same understanding of what is being attempted in the classroom. But equally so, this is true in the areas of the development of curriculum, the goals for instruction, and the design of assessments. The reporting of assessment results to parents and other actors in the education system should reflect these same understandings, as should the evaluations of effectiveness built into all systems. Each student should have an equivalent opportunity to achieve the defined goals, and the allocation of resources should reflect those goals. While state standards should be the basis for coherence, and should serve to establish a target for coordination of action within the system, many state standards are so general that they do not provide sufficient guidance about what is expected. Thus, each teacher, student, and assessment developer is left to decide independently what it means to attain a standard—a situation that can lead to curriculum, instruction, and assessment working at cross purposes. Better specified standards, especially standards that are organized according to the coherences mentioned above, can assist states in achieving coherence among curriculum, instruction, and assessment.

In a recent volume (Wilson, 2004a), where leading theorists and practitioners in educational assessment were asked to discuss the relationship between large-scale and classroom assessment, the commentary authors reached some common ground on a number of points (although there were still quite a few on which they differed):

(a) the current use of standardized tests for large-scale assessment has a negative effect on instruction and curriculum in the classroom;

(b) while seeing a legitimate role for standardized tests, they saw an essential role for other sorts of tests in expanding the range and depth of assessments, and hence reflecting essential outcomes of the curriculum;

(c) but, they saw a danger in tying the two levels, classroom and large-scale, *too* closely together, say, by using the same items for both.

Thus, coherence is seen as being somewhat complex--coherence doesn't mean "the same"--classroom assessment is not just large-scale assessment that takes place in classrooms. There must be a legitimate place for classroom assessment alongside large-scale assessment, and the two must be different. But how should we achieve this? The purpose of this paper, then, is to sketch out an answer to this question. There will be no claim that this answer is the only possible one, but many of its features will be required by any solution to the problem.

The next several sections lay out the three different modes of coherence, as described above, so that the argument can be clearly articulated. The main focus is on the topic of developmental coherence, as that is the core of the educational task: Hence, quite a bit of space is devoted to discussing the concepts of learning progressions and the closely-related topic of the BEAR Assessment System, which is used as a practical route to developing a learning progression. The paper concludes with a discussion of the importance of the position taken, and some reflections on what is needed for educational assessment to be a help and not a hindrance for educational progress.

Vertical coherence: systematic or by threat?

Wildy (2004) makes the salient point that a unifying idea in educational assessment is “that it is possible to bring classroom and large-scale assessments together conceptually in support of student learning.” In pointing this out, she is highlighting vertical coherence which was argued for by the NRC report *Knowing What Students Know*: “the conceptual base or models of student learning underlying the various external and classroom assessments within a system should be compatible” (NRC, 2001, p. 255). We will refer to this form of vertical coherence as *systemic* coherence—where there is a high degree of consistency between the underlying frameworks for the large-scale and classroom assessments—and below will make this meaning clearer by contrasting it with an alternative form of vertical coherence: coherence by *threat.* A summary of the relationship between the different types of vertical coherence that will be discussed in this section is provided in Figure 2.

==================
Insert Figure 2 about here
==================

Systemic coherence is not the only form of vertical coherence that one might find in an accountability system. Indeed, a more basic and probably more common form of vertical coherence is what we will refer to as*coherence by threat*—that is, where the large-scale assessment is used as a driving and constraining force, strait-jacketing the classroom instruction and curriculum to adhere to a specific curriculum, and hence meaning that classroom assessments are either (a) parallel to the large scale assessments, or (b) irrelevant for accountability. Essentially, classroom assessments become just a small component of vertical coherence in this case--the strait-jacket is directly imposed on classroom curriculum and instruction via the large-scale assessments, and classroom assessments are expected to follow suit. The import of this view was once made clear by a member of the State Board of Education for California, who said that the tests used for accountability don't have to be particularly good tests, they just have to serve their purpose—which is to ensure that teachers teach the standards as specified by the state (and, of course, expressed through the State tests)! It is interesting to note that, in the case where large-scale assessments are used in this way, vertical coherence need not be systemic in the way described above, it need only convey the right sort of *threat* to the classroom--hence the label. Note that this threat will not be effective unless the large-scale test is also used in a way that makes the classroom in at least some way (either directly through the teacher, or indirectly through the school) subject to the sanctions of an accountability system. For example, a standardized test can be used for accountability even if it does not actually measure *any* standard, yet, because it is "based on" the standards (i.e., items can individually be mapped to specific standards), the threat of inclusion of each and every standard will control the curriculum in the classroom (Suter, 2004).

In the current US context, where the Federal No Child Left Behind legislation has accelerated state assessments in mathematics and reading into every classroom in every state (with science following close behind), threat coherence is very much a reality. Although there are notable exceptions (see the examples in NRC reports--2003 and 2006), a common way that states have been complying with the legislation is to adapt or develop large-scale standardized tests (Erpenbach, Forte-Fast, & Potts, 2003). Typically, these sample quite lightly from among those state standards that are easily assessable with multiple-choice items. Regardless of the quality of these multiple choice items, (a) there are important aspects of school curricula that are not adequately assessable by multiple choice tests (Black and Wiliam, 2004; Thier, 2004); and (b) the sheer number of standards addressed in any given test (upwards of 100 in some states), ensure that the results cannot be used to gauge student accomplishment in a way that is useful in the classroom or the school for educational planning—all that is knowable on the basis of the tests is that students are doing better, worse, or about as well as, their peers on very broad areas such as "math" or "reading".

Conceptual, information and item coherence. Within the concept of systemic coherence as described above, one can think of several different degrees of specificity of the vertical coherence. At the weakest level, which we will refer to as *conceptual coherence*, the assessments at the classroom and large-scale levels share a common underlying framework. An example would be the "progress variables" described by Forster and Masters (2004). At the other extreme is a third level, which we will refer to as *item coherence*, where the actual tests and items (or clones of the items) used at one level would also be used at the other. For example, one type of item coherence would be to take the items used in large-scale assessments and use them for practice in the classroom. Another example, widely-referred to as "benchmark testing" would be where slightly-altered versions of the large-scale items were used periodically to check up on students throughout the year. Note that these different levels of systemic coherence are hierarchical—so that an implementation that satisfied the higher levels would necessarily satisfy the lower levels; hence, an example of item coherence would also necessarily be a, very strong, example of conceptual coherence.

A middle level of vertical coherence between these two extremes would be one that (a) shared a common framework, and (b) shared information between the classroom and the large-scale levels, but (c) did not necessarily use the same tests or results at the two levels—we will call this *information coherence*. An example of information coherence has been described by Wilson (2004b)—there it was called a “community of judgment.” The key aspect of the approach is that different sources of information are seen as having particular strengths, and the system is then designed to take advantage of those strengths (and avoid weaknesses). For example, where instructional validity is seen as being important, assessments that are close to the classroom are used, but, in order to ensure comparability across classroom contexts, information on consistency and verifiability must also be available, and that might come from large-scale tests, or from judgments by outside experts. Some specific Australian examples of how information has been shared across levels are described below in the next paragraph.

For example, several varieties of information-coherent assessment systems have been used at various times in different Australian states (Wilson, 1992a). One is typified by the process of statistical moderation of teachers’ in-class judgments—under this model, teachers in each school judge their students on locally-developed assessments, using centrally-developed guidelines. The students are also each given a centrally-developed test on the same material (i.e., the test is developed using the same framework as the local assessments, but perhaps using a more restricted item format). The test results are used to statistically moderate the school distributions (i.e., the mean and variance for the teacher’s assessments for each school are linearly transformed to be consistent with the mean and variance for their school on the test). But the within-school values are kept the same as the teacher assessments. Thus, the relative scores of students within schools are determined by the teachers’ assessments, but the between-school variability is determined by the test (McGaw, 1977; McGaw, Warry & McBryde, 1975). A different variant of such a system has been employed in the Australian state of Queensland, where the effect of the statistical moderation has been lightened so that any differences between the teachers’ assessments and the test are used ONLY to flag inconsistencies that are then followed up by an audit (NRC, 2003). Another variant on such an information-sharing system could use a work-sampling arrangement, with “experts” judging the teachers’ assessments of samples of student work, which could then be used as a basis for statistical moderation, or monitoring, as above.

Concern over the problems of item coherence have been raised (Black and Wiliam, 2004; Suter, 2004). LeMahieu and Reilly (2004) noted: "Each case that seeks to expand the utility of classroom assessment in service to accountability does so with the well-intended goal of rendering accountability contingent upon assessments that are coherent with classroom goals and practices. However a certain cautionary tone in the authors' writing betrays a reluctance to commit to the position that one assessment can serve both purposes wholly." Smithson (2004) explicitly argues against item coherence, citing problems with (a) feasibility and (b) desirability. Specifically, he doubts whether teacher assessments on performance assessments can be trusted in cases (a) where individual assessment purposes such as grade promotion and graduation are concerned, and (b) where accountability purposes such as rewards and sanctions for school districts, schools, and teachers themselves are concerned.

Going further, Shepard (2004) sees a further problem with coherence *per se*. She argues that coherence at the conceptual level or above requires a shared curriculum, and she sees this curriculum as being much more constraining than current conceptions of test alignment would indicate. In particular, she sees a framework such as a construct map (see below) as being more specific than current versions of state curricula as expressed through standards and potentially leading to an undesirable amount of curricular uniformity.

Of course, these are debatable matters. For some it might seem a bit odd, for example, to doubt (as does Smithson), the validity of teachers' judgments of performance assessments in a high-stakes context, yet to rely on teachers to decide students' grades, that are routinely used for making high-stakes decisions about those same individual students. And whether or not adoption of a framework based on information coherence (such as one based on the "construct maps" described below) was unduly restrictive would depend very much on the generality of those construct maps. The doubts discussed above are important ones, however, and would need to be addressed in any planned implementation of a system based on either information coherence or item coherence.

The application of these concepts of coherence to the current context under the No Child Left Behind legislation leads one to ask what varieties of vertical coherence can be found in different states. The legislation itself does not opt for any particular one of the levels of vertical coherence described above—the "alignment" of tests to state standards can be satisfied by any of the levels of coherence, from coherence by threat to item coherence. While most states are acting at the level of coherence by threat, others are attempting to achieve coherence at higher levels (see NRC, 2003). Unfortunately, given the scale of the systems that states are required to institute, and given the constant problems of funding, the most likely outcome is that states will satisfy the NCLB requirements in the cheapest and simplest way possible, which will likely mean that threat coherence will remain the norm unless serious efforts are made to develop alternative models and strategies, and to fund initial implementations of such alternatives.

Thus, in summary, one can say that the accountability and assessment systems currently predominant across the states under NCLB are based on an approach that will not foster the sorts of instruction and learning that are desperately needed for the education of citizens in the 21st century (NCR, 2007). However, the next section seeks to show that alternative approaches are available that can make positive use of the strength of the relationship among assessment, instruction, and learning, and thus be a catalyst for educational improvements that go beyond the accountancy aspects of accountability.

Developmental Coherence: Learning progressions

Note that the argument made here is not against the use of tests for monitoring—it is instead directed at the problems of bringing a useful degree of interpretability to assessments. The importance of this point goes well beyond assessments, however, as the assessment problem really arises from a curriculum problem. This matter has been brought to national attention by William Schmidt and his colleagues, who, in their analyses of many curricula from across the world, have developed an apt description for US curricula: “a mile wide and an inch deep” (Schmidt, McKnight & Raizen, 1997). They found that, compared to the (specifically mathematics and science) curricula in other countries, US curricula do not develop deep understanding of subject matter, but instead tend to spread their attention across a very broad set of domains, doubtless to satisfy as many professional and political groups as possible. And typical standardized tests reflect this curricular reality (as they must, in order to survive in the marketplace). This then brings us to the topic of developmental coherence. We 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 BEAR Assessment System.

At a recent meeting of researchers working on the topic of leaning progressions, the following broad description of learning progressions was suggested by a group consensus:

> 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. (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. As well, 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. Thorndike, 1904; Thurstone scaling; Thurstone, 1925; Guttman scaling; Guttman,1944; Bloom’s taxonomy; Bloom, 1956), the history of the specific term “learning progression” in the context of science education is a relatively brief one (CCII, 2009), starting with the publication of an NRC report (NRC, 2006). That report was focused on assessment in K-12 education, and hence the connections to assessment have been there right from the start. Nevertheless, given the brief time-span since then, there is not a great deal of extant literature regarding the relationship between the two, although this may well change in the near future. A second NRC report (NRC, 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*. 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 BEAR Assessment System (Wilson, 2005 Wilson & Sloane, 2000). In this paper we will draw on the latter as the core set of assessment perspectives and practices to relate to learning progressions.

The BEAR Assessment System (BAS)

The BEAR Assessment System is based on the idea that good assessment addresses the need for sound measurement through four principles: (1) a developmental perspective, (2) a match between instruction and assessment, (3) the generating of quality evidence, and (4) management by instructors to allow appropriate feedback, feed forward and follow-up. These four principles, plus four building blocks that embody them are shown in Figure 3. Below we take up each of these principles and building blocks in turn. See Wilson (2005) for a detailed account of an instrument development process that works through these steps.

=====================

Insert Figure 3 about here

=====================

Principle 1: A 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. Establishing appropriate criteria for taking a developmental perspective has been a challenge 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. Taxonomies such as Bloom's Taxonomy of Educational Objectives (Bloom, 1956), Haladyna's Cognitive Operations Dimensions (Haladyna, 1994) and the Structure of the Observed Learning Outcome (SOLO) Taxonomy (Biggs & 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 educational 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, the curriculum developers, and the assessment developers involved in the process of creating the assessments.

<u>Building Block 1: Construct Maps</u> Construct maps (Wilson, 2005) embody this first of the four principles: that of a developmental perspective on assessment of student achievement and growth. A construct map is a well thought out and researched ordering of qualitatively different levels of performance focusing on one characteristic. For a construct map to do its job well, it must resolve the tension between the following two characteristics: (a) it defines what is to be measured or assessed in terms general enough to be interpretable within a curriculum and potentially across curricula, but (b) it is specific enough to guide the development of the other components. When instructional practices are linked to the construct map, then the construct map also indicates the aims of the teaching. Construct maps are one model of how assessments can be integrated with instruction and accountability. They provide a way for large scale assessments to be linked in a principled way to what students are learning in classrooms, while at least having the potential to remain independent of the content of a specific curriculum.

This approach assumes that, within a given curriculum, student performance on curricular variables can be traced over the course of the curriculum, 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 certain 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 (i.e., a in terms of learning performances), are necessary to establish the construct validity of an assessment system. Such an approach provides a means for ensuring both developmental and horizontal coherence. If the progress variables around which a particular curriculum and its accompanying assessments are based are part of the state standards, vertical coherence is provided for as well.

The idea of using construct maps as the basis for assessments 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 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 variables that do not overlap with typical curricula, but the remainder will form a fairly stable set of variables that will be common across many curricula.

Construct maps are derived in part from research into the underlying cognitive structure of the domain and in part from professional judgments 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, consider the following example.

The example explored in this brief introduction is a test of science knowledge, focusing in particular on earth science knowledge in the area of “Earth in the Solar System" (ESS). The items in this test are distinctive, as they are Ordered Multiple Choice (OMC) items, which attempt to make use of the cognitive differences built into the options to make for more valid and reliable measurement (Briggs, Alonzo, Schwab & Wilson, 2006). The standards and benchmarks for “Earth in the Solar System” appear in Appendix A of the Briggs et al article (2006). According to these standards and the underlying research literature, by the 8th grade, students are expected to understand three different phenomena within the ESS domain: (1) the day/night cycle, (2) the phases of the Moon, and (3) the seasons -- in terms of the motion of objects in the Solar System. A complete scientific understanding of these three phenomena is the top level of our construct map. In order to define the lower levels of our construct map, the literature on student misconceptions with respect to ESS was reviewed by Briggs and his colleagues. Documented explanations of student misconceptions with respect to the day/night cycle, the phases of the Moon, and the seasons are displayed in Appendix A of the Briggs et al (2006) article.

The goal was to create a single continuum that could be used to describe typical students’ understanding of three phenomena within the ESS domain. In contrast, much of the existing literature documents students’ understandings about a particular ESS phenomena without connecting each understanding to their understandings about other related ESS phenomena. By examining student conceptions across the three phenomena and building on the progressions described by Vosniadou & Brewer (1994) and Baxter (1995), Briggs and his colleagues initially established a general outline of the construct map for student understanding of ESS. This general description helped them impose at least a partial order on the variety of student ideas represented in the literature. However, the levels were not fully defined until typical student thinking at each level could be specified. This typical student understanding is represented in the ESS construct map shown in Figure 4 (a) by general descriptions of what the student understands, and (b) by limitations to that thinking in the form of misconceptions, labeled as "common errors." Common errors used to define level 1 include explanations for day/night and the phases of the Moon involving something covering the Sun or Moon, respectively.

=====================
Insert Figure 4 about here
=====================

In addition to defining student understanding at each level of the continuum, the notion of common errors helps to clarify the difference between levels. Misconceptions, represented as common errors in one level, are resolved in the next level of the construct map. For example, students at level 3 think that it gets dark at night because the Earth goes around the Sun once a day—a common error for level 3—while students at level 4 no longer believe that the Earth orbits the Sun daily but rather understand that this occurs on an annual basis.

The top level of the ESS construct map represents the understanding expected of 8th graders in national standards documents. Because students' understanding of ESS develops throughout their schooling, it was important that the same continuum be used to describe the understandings of both 5th and 8th grade students. However, the top level is not expected of 5th graders; equally, we do not expect many 8th grade students to fall into the lowest levels of the continuum.

Principle 2: Match between Instruction and Assessment

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 how their use ensures that the framework for the assessments and the framework for the curriculum and instruction are one and the same.

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 and typical student responses are matched to certain levels within at least one construct map. This should be true for assessments at all levels of the system, from classroom formative assessments to state standardized assessments. This helps to ensure an appropriate level of vertical and horizontal coherence within the system.

Returning to the ESS example, the OMC items were written as a function of the underlying construct map, which is central to both the design and interpretation of the OMC items. Item prompts were determined by both the domain as defined in the construct map and canonical questions (i.e., those which are cited in standards documents and commonly used in research and assessment contexts). The ESS construct map focuses on students' understanding of the motion of objects in the Solar System and explanations for observable phenomena (e.g., the day/night cycle, the phases of the Moon, and the seasons) in terms of this motion. Therefore, the ESS OMC item prompts focused on students' understanding of the motion of objects in the Solar System and the associated observable phenomena. Distractors were written to represent (a) different levels of the construct map, based upon the description of both understandings and common errors expected of a student at a given level and (b) student responses that were observed from an open-ended version of the item. Two sample OMC items, showing the correspondence between response options and levels of the construct map are shown in Figure 5 Each item response option is linked to a specific level of the construct map. Thus, instead of gathering information solely related to student understanding of the specific context described in the question, OMC items allow us to link student answers to the larger ESS domain represented in the construct map. Taken together, a student's responses to a set of OMC items permit an estimate of the student's level of understanding, as well as providing diagnostic information about specific misconceptions.

=====================
Insert Figure 5 about here
=====================

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 behind the progress variables. Open-ended tasks, if used, must be quickly, readily, and reliably scorable.

<u>Building Block 3: The outcome space</u> The outcome space is the set of categorical 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 variables.

Principle 4: Evidence of High Quality Assessment

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure along 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 consistency 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 achievement variables, and (d) establish uniform levels of system functioning, in terms of quality control indices such as reliability.

<u>Building Block 4: Wright maps</u> Wright maps represent this principle of evidence of high quality. Wright maps are graphical and empirical representations of a construct map, showing how it unfolds or evolves in terms of increasingly sophisticated student performances.

Wright maps allow us to examine the quality of our assessments in a number of important ways. By examining the relative difficulty of the items, and the response levels within those items, we can make sure that these reflect our predictions (based on our original theory). By examining the relative fit of the items, we can identify those items that are not performing as required. By examining person fit, we can identify those persons for whom the particular assessment does not appear to be a good summary of that student's knowledge and skill. In addition, the models used to produce the Wright maps can provide us with traditional reliability coefficients, estimates of rater severity for items scored by raters, and the effects of various person demographics (e.g. gender, ethnicity) which may be a source of differential item functioning (DIF).

Broader Perspectives on the BAS

We typically use a multi-dimensional Rasch modeling approach to calibrate the maps for use in the BEAR Assessment System (see Adams, Wilson, & Wang (1997) and Briggs & Wilson (2001), 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. Later in this paper, we will use a somewhat different approach to modeling in order to integrate hypotheses about links among constructs within a learning progression. This will alter some of the statistical aspects of the modeling, but many of the types of analyses and results will remain similar

In this brief summary, 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 construct maps to provide a common conceptual framework across curricula. Construct maps 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. These construct maps 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.

An example of how this might play out in a given curricular unit is provided by a second example, drawn from the Buoyancy unit of the FAST Curriculum (Kennedy and Draney, 2007), a focus of study by members of the Center for Assessment and Evaluation of Student Learning (CAESL), a Center for Teaching and Learning funded by the National Science Foundation. The topics addressed in this unit were density and its relationship to floating and sinking. Students learned first about mass, then volume, then density, and finally relative density, as illustrated in Figure 6. At key points during the curriculum, formative assessments, referred to in the figure as "reflective lessons", were given to the students. These included "predict-observe-explain" activities, as well as answering as completely possible the question "Why do things sink and float?" Their answers to such open-ended questions were scored using the Scoring Guide shown in Figure 7 (drawn directly from a construct map). In addition, the pretest and posttest for this unit included multiple-choice questions on the same topic, which could easily be included in classroom summative assessments, as well as in standardized tests.

=======================

Insert Figure 6 and 7 about here
========================

With the assessments to be used across curricula structured by construct maps, the problem of item development is lessened – ideas and contexts for assessment tasks may be adapted to serve multiple curricula that share construct maps. 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.

Mapping out a learning progression using construct maps

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

In order to illustrate certain aspects of the relationship between learning progressions and assessment, we will use a visual metaphor that superimposes images of construct maps on an image of a learning progression. This image of the learning progression is shown in Figure 8, 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 a someone (a science educator, a science education researcher, an assessment developer?) who is thinking about student thinking.

=====================
Insert Figure 8 about here
=====================

The relationship between the construct maps that make up the learning progression may be quite complex (See Wilson (2008) and Wilson (in press) for examples of other relationships between the construct maps and the learning progression.) One straightforward way to see the relationship of construct map to learning progression is to see the learning progression as composed of a set of construct maps, each comprising a "dimension" of the learning progression, and 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 3-dimensional space originating from a common source, as is common in geometric interpretations of psychometric models. Here the angle between the arrows is an indicator of the correlation between the dimensions.

To illustrate this assessment structure, we 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 9, superimposed on 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, and that may indeed be the case, conceptually, but 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.

=====================

Insert Figure 9 about here

=====================

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 10. Here, the attainment of levels of a construct would be seen as being dependent on 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, is shown in Figure 11 (Wilson & Black, 2007). 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.

=====================

Insert Figures 10 & 11 about here

=====================

A more complicated way of seeing such a possibility is shown in Figure 12, where there are links hypothesized that are between specific levels of one construct, and specific levels of other constructs (rather than the "top to bottom" relationships shown in Figure 13).

=====================

Insert Figures 12 & 13 about here

=====================

Horizontal Coherence: Learning Performances[1]

A concept that can be useful in considering the horizontal coherence among the curriculum, instruction and assessment 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" (NRC, 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, measured. 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: Smith, Wiser, Anderson, Krajcik, and Coppola (2004) summarized a set of observable performances that could provide indicators of understanding in science (see Figure 14).

=====================

[1] This section is adapted from NRC (2006), pp 91-94.

Insert Figure 14 about here

====================

As a concrete example, take the following standard that is adapted from *Benchmarks for Science Literacy* (AAAS, 1993, p. 124) about differential survival:

> [The student will understand that] *Individual organisms with certain traits are more likely than others to survive and have offspring.*

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) amplify this single standard as three related learning performances:

- Students *identify and represent mathematically* the variation on a trait in a population.
- Students *hypothesize* the function a trait may serve and *explain* how some variations of the trait are advantageous in the environment.
- 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) advance 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) that is being assessed. 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 could have been left out of the test.

In the context of the BEAR Assessment System, 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 video-tapes 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).

Discussion

This paper has addressed an issue that is prominent at this time--how to develop core ideas in science education. One very important aspect of this is how to relate large-scale assessments to the classroom situation. One might almost say that the wish of every large-scale testing program is to have the results of the large-scale testing be useful to teachers in the classroom. This is something that is asked for by State Testing Directors, and promised by testing companies. But, it is seldom or never that sufficient information is available in large-scale assessments for such micro-scale usage, although some testing programs do indeed provide the raw information (invariably without supplying standard errors) to users. Even if there were sufficient information available, there are strong doubts that positive outcomes would result from in-classroom usage of large-scale assessments due to the nature of the current mainstream approach to large-scale testing. In order that large-scale and classroom assessments can mutually support one another, the recommendation of this paper is that both the large-scale and the classroom assessments must be constructed to be coherent in an educational sense, and that should, of course, be based upon a developmentally-coherent approach such as through learning progressions.

In this context, the BEAR Assessment System was described as a process that can establish the coherence necessary to allow the micro and the macro to function coherently together, the coherence necessary to follow student development over a long period, and the coherence necessary to relate the assessments usefully to the instruction. As an example, the paper described an assessment application in the area of Earth Sciences that illustrates these ideas, and shows some of the features of a coherent system. Implementation of a BEAR Assessment System is not a minor matter, however. It requires a deeper analysis of the relationship between student learning and the curriculum and instructional practices than is commonly the case in assessment development. It also requires a readiness to revise curriculum (i.e., "standards") based on empirical evidence available from the results of the assessments. While this latter is logically clear to most people, the political environment of accountability systems and debates about standards can make it almost impossible to achieve. To attain the wished-for effects on educational achievement, much work is also needed in developing materials to help teachers who will be using the information from the assessments, and also to provide professional development for those same teachers. In spite of this required effort, a number of BAS-inspired curriculum-based assessment systems are now taking shape (e.g., FOSS in elementary science and Living by Chemistry at the high school level), and more are under development[2].

In translating into the policy realm the perspectives, developmental work processes, and products that have been delineated in this paper, one must expand one's viewpoint somewhat.

First, as has been noted a few times in the text above, the arguments that have been advanced will sometimes be seen as outside of the policy decision-making process. This can be either because the policy-makers are not willing to cede a place for research and development and evidence-based arguments in their decisions, or (and this is sometimes just a disguised version of the previous case), the pressures of expediency are enjoined to eliminate the need to do so. This tendency must be resisted at all costs. One perspective we use is to compare the self-perceived roles of educational policy-makers with the roles taken on by policy-makers in other areas where public decisions must be made, for example in the area of highway building. Here, policy-making bodies must make ultimate decisions about very general issues concerning which highways to build, and whose interests to satisfy first, etc.. But, although such bodies would likely include a few specialists in highway design and construction, the bodies themselves understand that there are technical and professional dimensions to the decisions that they must make, and that they must seek out and use the best technical advice when they make those decisions. Unfortunately, this respect for relevant expertise is much absent in the equivalent policy decision-making bodies in the field of education (e.g., State Boards of Education, etc.). In education, presumably because all of the policy-makers have been educated, policy-makers are prepared to assume that they do indeed have the technical and professional knowledge to make detailed decisions, as well as the broader policy issues, which are their appropriate bailiwick. It is certainly true that policy-makers in the area of highways drive over highways, but they are sensible enough to understand that that does not mean that they are thus appropriately-skilled to

[2] See the BEAR Center website at:http://bearcenter.berkeley.edu/research.php

be deciding on the composition of technical issues such as the surfacing ingredients. This common-sense understanding of one's limits needs to be observed and understood by educational policy-makers.

Second, assuming that the issues discussed in the previous paragraph have been dealt with, it is important to appreciate that certain parts of the perspective presented above are designed to make a greater depth of policy argumentation and information accessible to educational policy-makers. This is particularly true of the learning progression concept, which has the possibility to provide a new level of discussion in the policy arena. With the business-as-usual standards approach, policy-makers have been left with a situation where they are confronted with a forest of standards to authorize. In some states, there are up to 90 or a 100 per grade, and, although they will typically be sub-divided into some sort of curriculum unit, such as "earth sciences," etc., these groupings are too coarse to help a teacher plan their instruction from week to week or even month to month and hence do not make it possible to make educationally-significant decisions. This is one of the potentials of learning progressions—policy-makers can have the "language" to make decisions about educationally-significant, but still manageable units of curriculum. They can decide to make some progressions "core" and some optional. They can decide to focus their curricula on particular aspects of science content and/or science process and practice. And they can shape student's growth over the years of the whole school curriculum. They can also ask educational researchers to develop new progressions, and evaluate the effects of old ones. They can even involve parents and teachers in these debates. This potential role for learning progressions is one very positive possible outcome of adopting this approach.

In conclusion, there are numerous ways that we can improve our process for setting, and assessing the ideas, both core and otherwise of the educational system that we have. In this paper, we have attempted to lay out some issues, and some possible approaches that we believe will be helpful in this process.

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

American Association for the Advancement of Science. (1993)*Benchmarks for Science Literacy*. New York: Oxford University Press.

Baxter, J. (1995). Children's understanding of astronomy and the earth sciences. In S.M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 155-177). Mahwah, NJ: Lawrence Erlbaum Associates.

Biggs, J. B., & Collis, K. F. (1982).*Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003).*Assessment for learning*. London: Open University Press.

Black, P, & Wiliam, D (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.),*Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II* (pp. 20-50). Chicago: National Society for the Study of Education.

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

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Briggs, D. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Butter, R., De Boeck, P., & Verhelst, N. D. (1998). An item response model with internal restrictions on item difficulty.*Psychometrika, 63,* 1-17.

Center for Continuous Instructional Improvement (CCII). (2009)*Report of the CCII Panel on Learning Progressions in Science.* CPRE Research Report, Columbia University, New York.

De Boeck, P. (2008).*Random item IRT models.* Presidential address presented at the annual meeting of the Psychometric Society, Durham, NH.

De Boeck, P. & Wilson, M. (Eds.). (2004).*Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Draney, K, & Wilson, M. (2007). Application of the Saltus model to stage-like data: Some applications and current developments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 119-130). New York: Springer.

Duncan, R.G., & Hmelo-Silver. C. (in press). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal for Research in Science Teaching.*

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB: Central issues arising from an examination of state Accountability Workbooks and U.S. Department Of Education reviews under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State School Officers.

Forster, M., & Masters, G. N. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II* (pp. 51-73). Chicago: National Society for the Study of Education.

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

Haladyna, T. M. (1994). Cognitive taxonomies. In T. M. Haladyna (Ed.), *Developing and validating multiple-choice test items* (pp. 104–110). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hewson, P. W. (1992). *Conceptual change in science teaching and teacher education*. Paper presented at a meeting by the National Center for Educational Research, Documentation and Assessment. Ministry of Education and Science, Madrid, Spain, 1992.

Hoskens, M. & Wilson, M. (1999). *ConstructMap* [Computer program]. Berkeley, CA: Berkeley Evaluation and Assessment Research Center.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Kennedy, C. A. & Draney, K. (2007). Interpreting and Using Multidimensional Performance Data to Improve Learning in Science. In Liu, X. & Boone, W. J. (Eds.), Applications of Rasch Measurement in Science Education. Maple Grove, MN: JAM Press.

Kennedy, C.A., Wilson, M., & Draney, K. (2006). *ConstructMap* [Computer program]. Berkeley, California: Berkeley Evaluation and Assessment Center, University of California.

Lee, Y-S., & Wilson, M. (2009). An extension of the MIRID model for polytomous responses and Random Effects. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

LeMahieu, P. G., & Reilly, E. C. *(2004).* Systems of coherence and resonance: Assessment for education and assessment of education. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society*

Formatted: German Germany

for the Study of Education, Part II (pp. 189-202). Chicago: National Society for the Study of Education.

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

Maris, E. (1999). Estimating multiple classification latent class models*Psychometrika, 64*, 187-212.

Masters, G. N., Adams, R. J., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.),*International Encyclopedia of Education: Research and Studies. Supplementary Volume 2.* Oxford: Pergamon Press. (pp. 628-634)

Masters, G., and Forster, M. (1996).*Progress maps. Assessment resource kit.*Victoria,

Australia: Commonwealth of Australia.

McGaw, B. (1977). The use of rescaled teacher assessments in the admission of students to tertiary study.*Australian Journal of Education, 21* (3), 209-225.

McGaw, B., Warry, R., & McBryde, B. (1975)*The Queensland grade 12 study report No. 2, validation of aptitude measures for the rescaling of school assessments.* Brisbane: Research Branch, Department of Education.

Minstrell, J. (1998, October). *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.

Mohan, L., Chen, J., & Anderson, C. W. (2008)*Developing a K-12 learning progression for carbon cycling in socio-ecological systems*. Center for Curriculum Materials in Science Research Report, Michigan State University (Downloaded from http://edr1.educ.msu.edu/EnvironmentalLit/publicsite/html/carbon.html)

Muthén, L.K. and Muthén, B.O. (1998-2007).*Mplus User's Guide.* Fifth Edition. Los Angeles, CA: Muthén & Muthén.

National Research Council (NRC). (2001*) Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, J. Pellegrino, N. Chudowsky and R. Glaser (Eds.), Division on Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

National Research Council (NRC). (2003).*Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment.* Workshop report. Washington, DC: National Academies Press.

National Research Council (NRC). (2006).*Systems for state science assessment*, Committee on Test Design for K-12 Science Achievement, M. Wilson & M. Bertenthal, (Eds.), Division on Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

National Research Council. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.). Washington, D.C.: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115, Stat. 1425 (2002).

Popham, W. J. (2003). *Crafting curricula aims for instructionally supportive assessment*. Available: http://education.umn.edu/nceo/Presentations/CraftingCurricula.pdf

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change *Science Education, 66*(2), 211-227.

Programme for International Student Assessment. (2005) *Learning for tomorrow's world: First results from PISA 2003.* Paris: Organisation for Economic Co-operation and Development.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001). *GLLAMM Manual. Technical Report 2001*. Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London.

Reiser, B.J., Krajcik, J., Moje, E., and Marx, R. (2003). *Design strategies for developing science instructional materials.* Paper presented at the National Association for Research in Science Teaching Annual Meeting, March, Philadelphia, PA.

Formatted: German Germany

Resnick, L.B. (1995). From aptitude to effort: A new foundation for our

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 37–76). Boston: Kluwer.

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

Roussos, L.A., DiBello, L.V., Stout, W., Hartz, S.M., Henson, R.A., & Templin, J.L. (2007). The Fusion model skills diagnostic system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education*. Cambridge University Press, Cambridge, UK.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997) *A splintered vision: An investigation of U.S. science and mathematics education*. New York: Kluwer Academic Publishers.

Sheperd, L. A. (2004). Curricular coherence in assessment design. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of* *the National Society for the Study of Education, Part II* (pp. 239-249). Chicago: National Society for the Study of Education.

Smith, C., Wiser, M., Anderson, C.W., Krajcik, J., and Coppola, B. (2004). *Implications of research on children's learning for assessment: matter and atomic molecular theory*. Commissioned paper prepared for the National Research Council's Committee on Test Design for K–12 Science Achievement,Washington, DC.

Smithson, J. (2004). Converging paths: Common themes in making assessments useful to teachers and systems. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II* (pp. 209-216). Chicago: National Society for the Study of Education.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Suter, L. (2004). Tools for two masters: Classroom assessment and school system assessment. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II* (pp. 169-182). Chicago: National Society for the Study of Education.

Thier, H. D. (2004). Structuring successful collaborations between developers and assessment specialists. In M. Wilson (Ed.), Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II *(pp. 250-263). Chicago: National Society for the Study of Education.*

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: The Science Press.

Vermunt J. K., & Magidson J. (2008). *Latent Gold 4.5 user's guide*. Belmont, MA: Statistical Innovations Inc.

Vosniadou, S., & Brewer, W.F. (1994). Mental models of the day/night cycle. *Cognitive Science*, 18, 123-183.

Wildy, H.(2004). The data club: Helping schools use accountability data. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II* (pp. 155-168). Chicago: National Society for the Study of Education.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105(*2), 276-289.

Wilson, M. (1992a). *The integration of school-based assessments into a state-wide assessment system: Historical perspectives and contemporary issues*. BEAR Center: University of California, Berkeley: BEAR Research Report (Re-issued May 2000).

Wilson, M. (1992b). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement. 16*(3), 309-325.

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

Wilson, M. (Ed.). (2004b). Assessment, accountability and the classroom: A community of judgment. In M. Wilson (Ed.), *Towards Coherence between Classroom Assessment and Accountability. The 103rd yearbook of the National Society for the Study of Education, Part II* (pp. 1-19). Chicago: National Society for the Study of Education.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Erlbaum.

Wilson, M. (in press). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching.*

Wilson, M. (2008, March). *Measuring progressions*. In A. C. Alonzo & A. W. Gotwals (Chairs), Diverse perspectives on the development, assessment, and validation of learning progressions in science. Symposium conducted at the annual meeting of the American Educational Research Association, New York. Retrieved March 30, 2008, from http://myweb.uiowa.edu/alonzo/aera2008.html

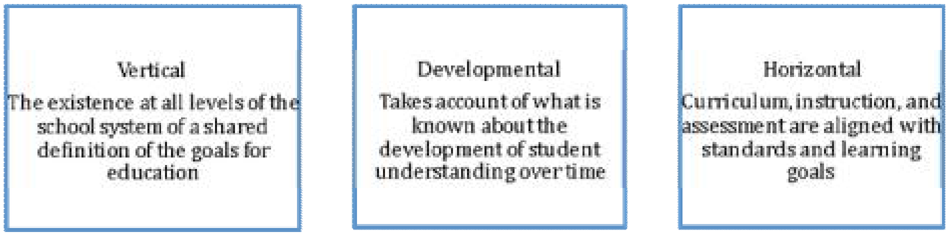
Wilson, M., & Black, P. (2007, March). *The idea of a learning progression as a core for both instruction and assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (pp. 325–332). Tokyo: Springer-Verlag.

Wilson, M., & Draney, K. (2005). *From principles to practice in assessment design: The BEAR Assessment System in a large-scale assessment context* (BEAR Research Report). Berkeley, California: University of California, Graduate School of Education.

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

Wolf, D. P., & Reardon, S. (1996). Access to excellence through new forms of student assessment. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities*. (Ninety-fifth Yearbook of the National Society for the Study of Education, part 1, pp. 52–83). Chicago: University of Chicago Press.



Vertical

The existence at all levels of the school system of a shared definition of the goals for education

Developmental

Takes account of what is known about the development of student understanding over time

Horizontal

Curriculum, instruction, and assessment are aligned with standards and learning goals

Figure 1: 3 Types of Coherence

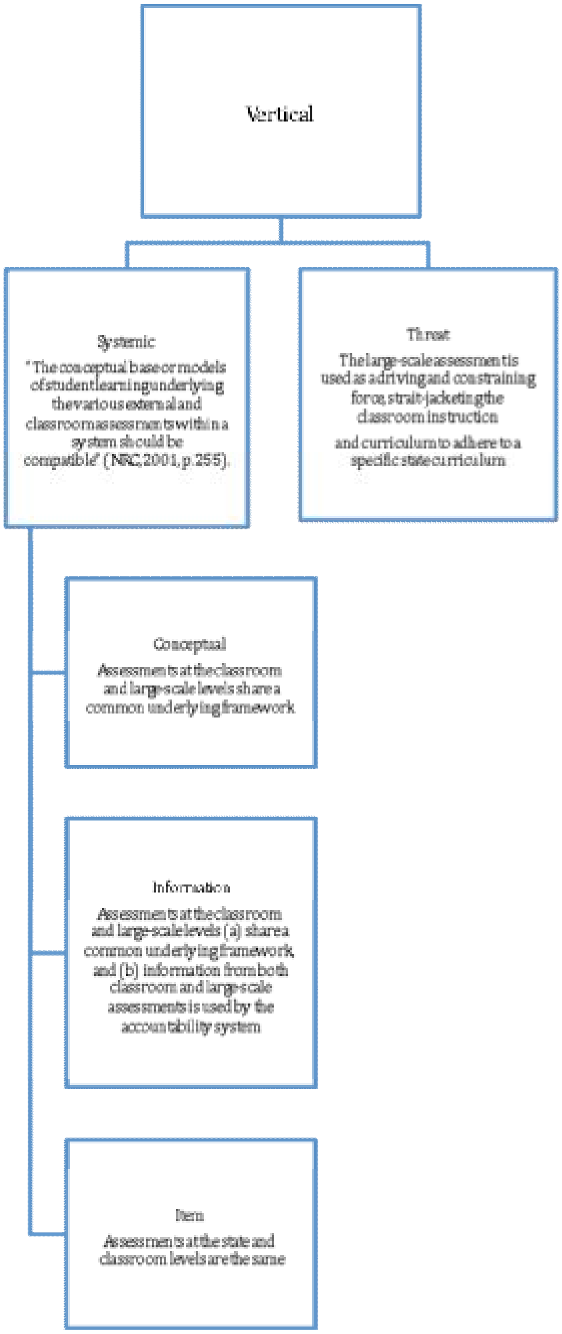

Vertical
Systemic
"The conceptual base or models of student learning underlying the various external and classroom assessments within a system should be compatible" (NRC, 2001, p.255).
Threat
The large-scale assessment is used as a driving and constraining force, strait-jacketing the classroom instruction
and curriculum to adhere to a specific state curriculum
Conceptual
Assessments at the classroom and large-scale levels share a common underlying framework
Information
Assessments at the classroom and large-scale levels (a) share a common underlying framework, and (b) information from both classroom and large-scale assessments is used by the accountability system
Item
Assessments at the state and classroom levels are the same

Figure 2: Types of vertical coherence

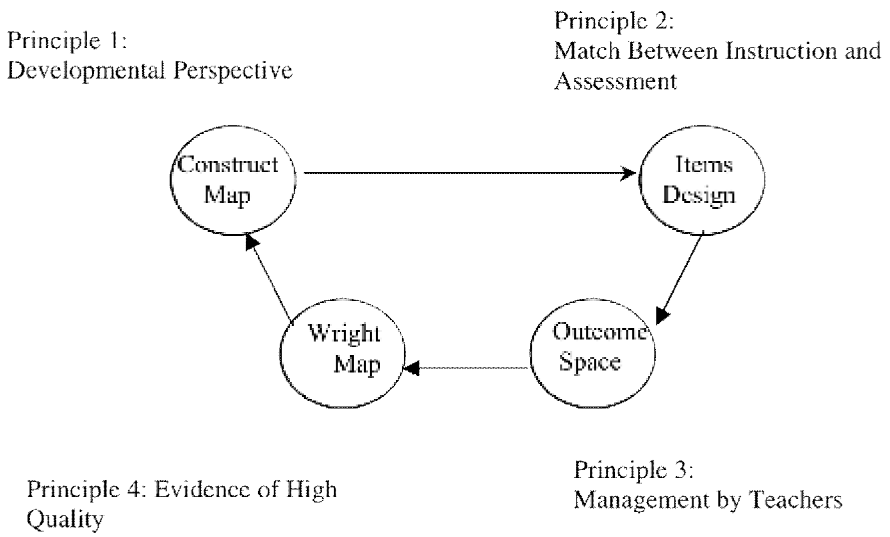


Figure 3. The principles and building blocks of the BEAR Assessment System.

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: • the day/night cycle • the phases of the Moon (including the illumination of the Moon by the Sun) • the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that • the Earth is both orbiting the Sun and rotating on its axis • the Earth orbits the Sun once per year • the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky • the Moon orbits the Earth once every 28 days, producing the phases of the Moon COMMON ERROR: Seasons are caused by the changing distance between the Earth and Sun. COMMON ERROR: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the Moon.
3	Student knows that: • the Earth orbits the Sun • the Moon orbits the Earth • the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. COMMON ERROR: It gets dark at night because the Earth goes around the Sun once a day.
2	Student recognizes that: • the Sun appears to move across the sky every day • the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis. COMMON ERROR: The Sun travels around the Earth. COMMON ERROR: It gets dark at night because the Sun goes around the Earth once a day. COMMON ERROR: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, "darkness") covers the Sun. COMMON ERROR: The phases of the Moon are caused by clouds covering the Moon. COMMON ERROR: The Sun goes below the Earth at night.
0	No evidence or off-track

Figure 4. Construct Map for Student Understanding of Earth in the Solar System

Item appropriate for fifth graders:

It is most likely colder at night because

A. the Earth is at the furthest point in its orbit around the Sun. Level 3

B. the Sun has traveled to the other side of the Earth. Level 2

C. the Sun is below the Earth and the Moon does not emit as much heat as the Sun. Level 1

D. the place where it is night on Earth is rotated away from the Sun. Level 4

© WestEd, 2002

Item appropriate for eight graders:

Which is the best explanation for why we experience different seasons (winter, summer, etc.) on Earth?

A. The Earth's orbit around the Sun makes us closer to the Sun in summer and farther away in winter. Level 4

B. The Earth's orbit around the Sun makes us face the Sun in the summer and away from the Sun in the winter. Level 3

C. The Earth's tilt causes the Sun to shine more directly in summer than in winter. Level 5

D. The Earth's tilt makes us closer to the Sun in summer than in winter. Level 4

© WestEd, 2002

Figure 5. Sample OMC items based upon Earth in the Solar System Construct map

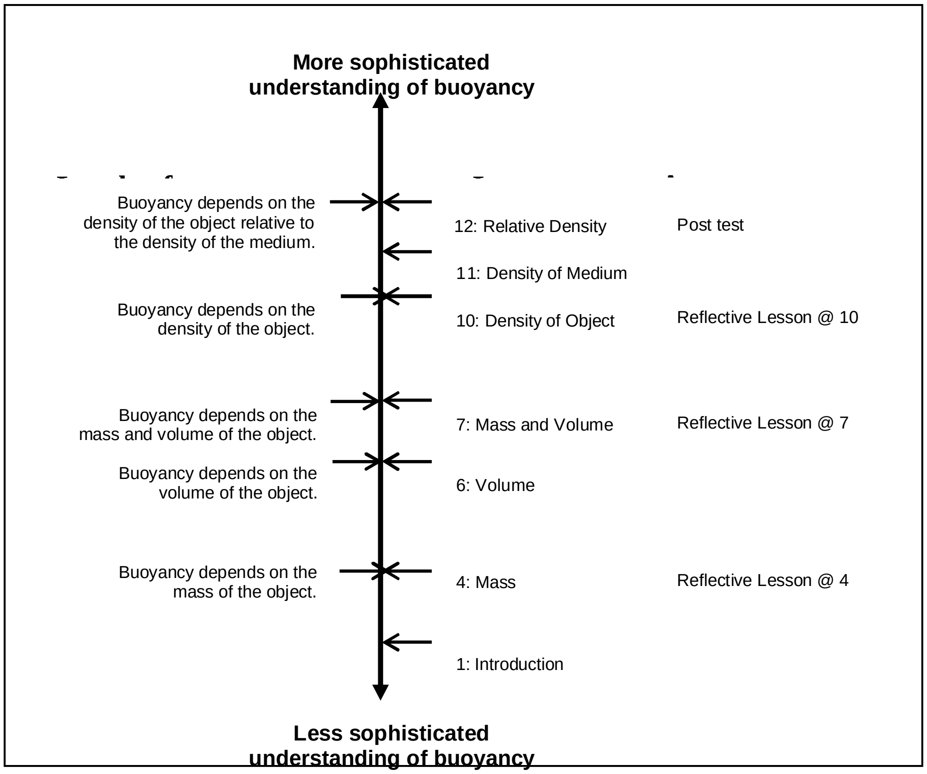

More sophisticated understanding of buoyancy
Buoyancy depends on the density of the object relative to the density of the medium.
12: Relative Density
Post test
11: Density of Medium
Buoyancy depends on the density of the object.
10: Density of Object
Reflective Lesson @ 10
Buoyancy depends on the mass and volume of the object.
7: Mass and Volume
Reflective Lesson @ 7
Buoyancy depends on the volume of the object.
6: Volume
Buoyancy depends on the mass of the object.
4: Mass
Reflective Lesson @ 4
1: Introduction
Less sophisticated understanding of buoyancy

Figure 6. Description of the “Why Things Sink and Float” unit. Shows the definitions and alignment of knowledge levels students are expected to exhibit, the knowledge levels targeted by the curriculum at several lessons, and the knowledge levels targeted by each assessment activity.

<table>
<tr><th colspan="5">Construct Map: Why Things Sink and Float</th></tr>
<tr><th colspan="2">Level</th><th colspan="2">What the Student Knows</th><th>Example Responses</th></tr>
<tr><td colspan="2">RD</td><td colspan="2"><b>Relative Density</b>
Student knows that floating depends on having less density than the medium.</td><td>“An object floats when its density is less than the density of the medium.”</td></tr>
<tr><td colspan="2">D</td><td colspan="2"><b>Density</b>
Student knows that floating depends on having a small density.</td><td>“An object floats when its density is small.”</td></tr>
<tr><td colspan="2">MV</td><td colspan="2"><b>Mass and Volume</b>
Student knows that floating depends on having a small mass and a large volume.</td><td>“An object floats when its mass is small and its volume is large.”</td></tr>
<tr><td>M</td><td>V</td><td><b>Mass</b>
Student knows that floating depends on having a small mass.</td><td><b>Volume</b>
Student knows that floating depends on having a large volume.</td><td>“An object floats when its mass is small.”
“An object floats when its volume is large.”</td></tr>
<tr><td colspan="2">MIS</td><td colspan="2"><b>Misconception</b>
Student thinks that floating depends on having a small size, being flat, filled with air, or having holes.</td><td>“An object floats when it is small.”
“An object floats when it is flat.”</td></tr>
<tr><td colspan="2">OT</td><td colspan="2"><b>Off Target</b>
Student does not attend to any property or feature to explain floating.</td><td>“I have no idea.”</td></tr>
</table>

Figure 7. Scoring Guides for the Why Things Sink and Float progress variable for the CAESL/FAST buoyancy curriculum

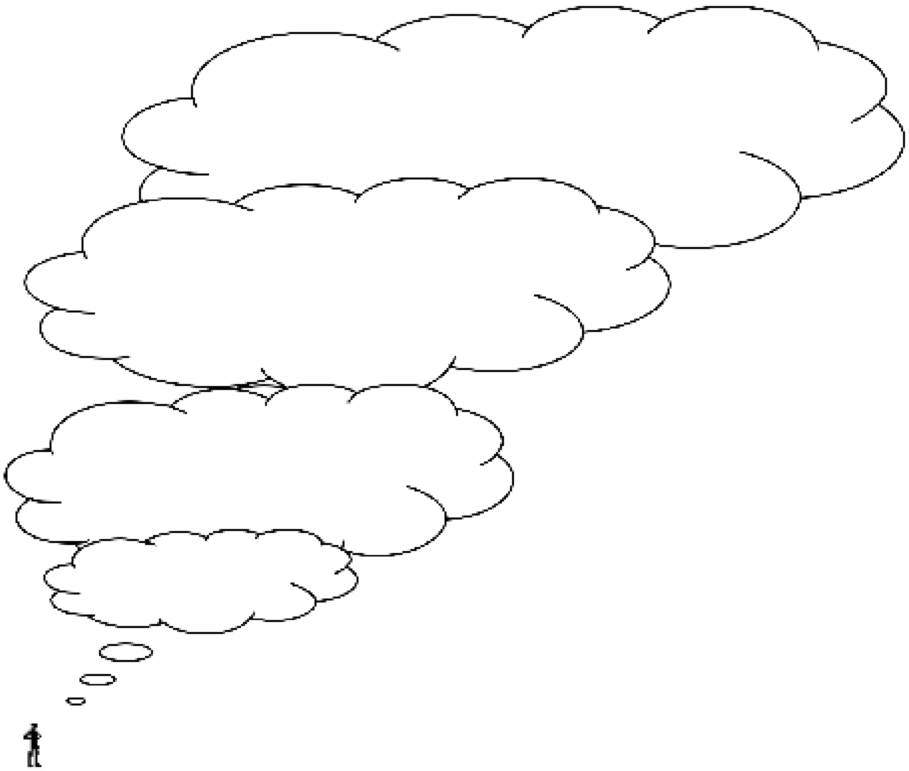


Figure 8. An image of a learning progression.

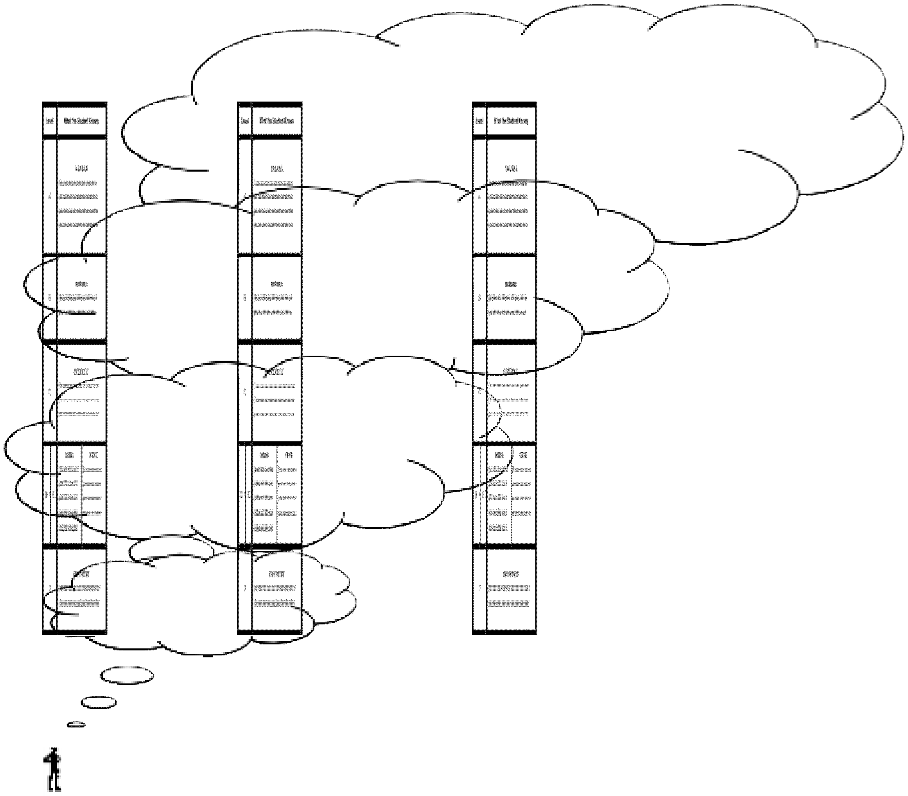


Figure 9. One possible relationship—the levels of the learning progression are levels of several construct maps.

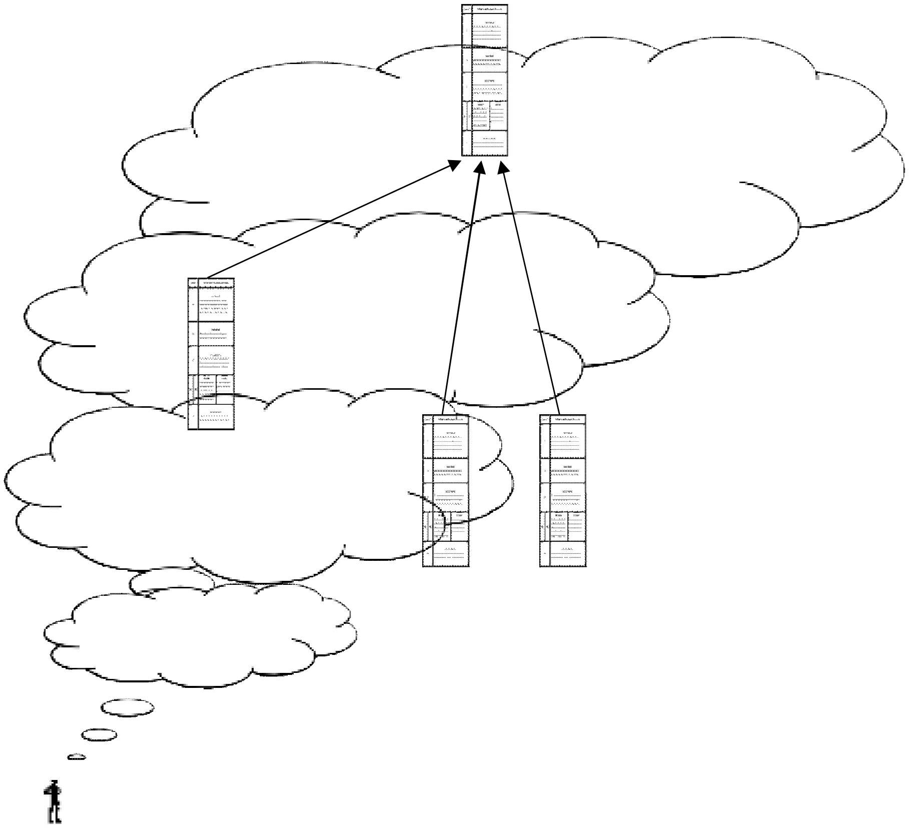


Figure 10. In this situation, there is a complicated dependency relationship between the construct maps in the learning progression.

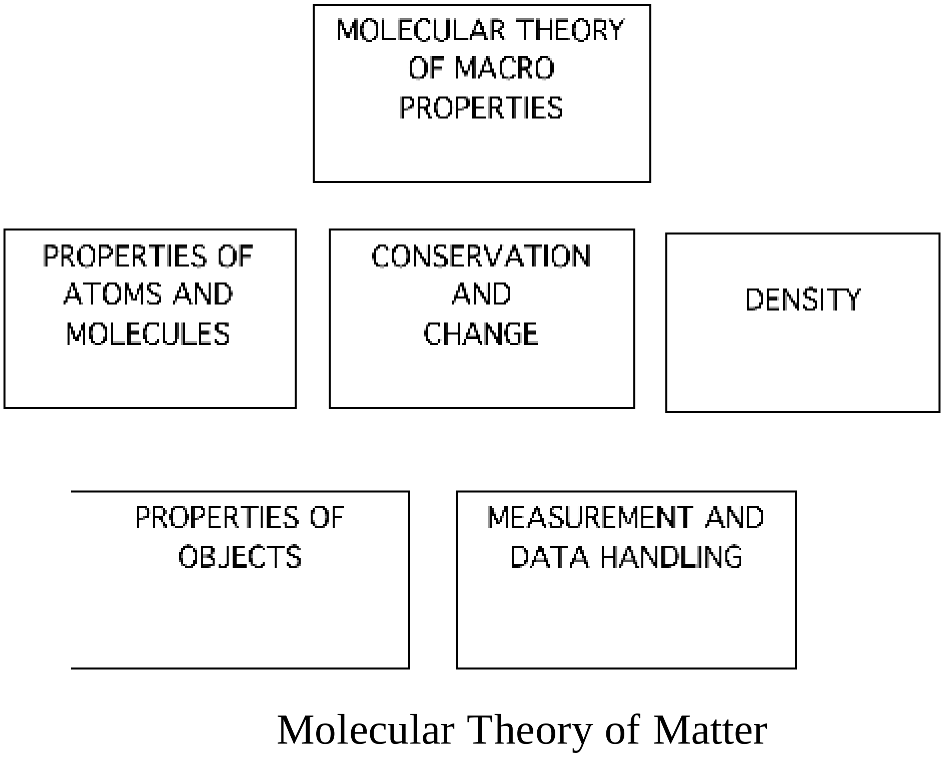

MOLECULAR THEORY OF MACRO PROPERTIES
PROPERTIES OF ATOMS AND MOLECULES
CONSERVATION AND CHANGE
DENSITY
PROPERTIES OF OBJECTS
MEASUREMENT AND DATA HANDLING
Molecular Theory of Matter

Figure 11. A set of constructs hypothesized to constitute a Molecular Theory of Matter.

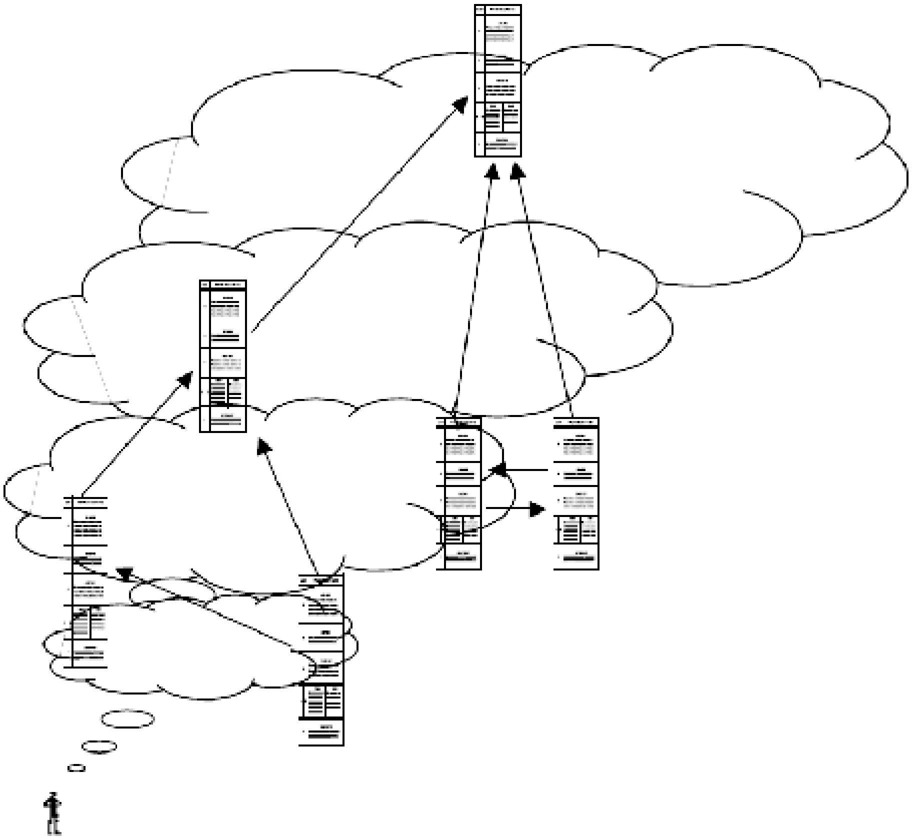


Figure 12. In this situation, the relationship between the construct maps is from level to level.

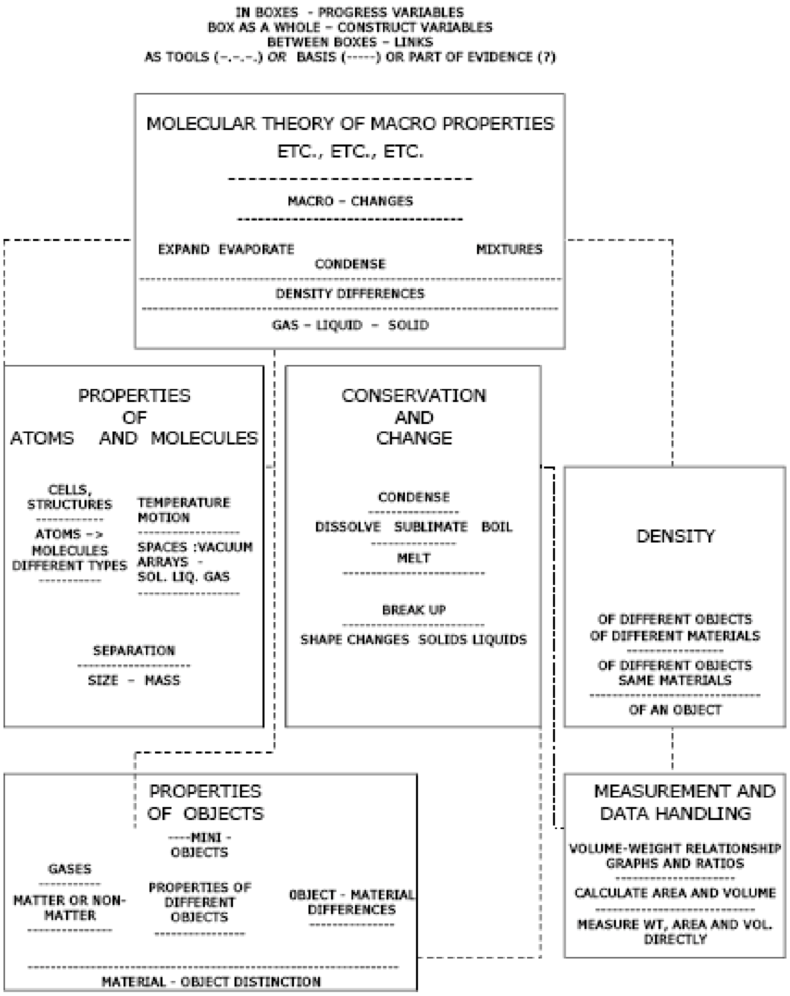

IN BOXES - PROGRESS VARIABLES
BOX AS A WHOLE – CONSTRUCT VARIABLES
BETWEEN BOXES – LINKS
AS TOOLS (–.–.–.) OR BASIS (-----) OR PART OF EVIDENCE (?)
MOLECULAR THEORY OF MACRO PROPERTIES
ETC., ETC., ETC.
MACRO – CHANGES
EXPAND EVAPORATE
CONDENSE
MIXTURES
DENSITY DIFFERENCES
GAS – LIQUID – SOLID
PROPERTIES OF ATOMS AND MOLECULES
CELLS, STRUCTURES
ATOMS –> MOLECULES DIFFERENT TYPES
TEMPERATURE MOTION
SPACES :VACUUM ARRAYS - SOL. LIQ. GAS
SEPARATION
SIZE – MASS
CONSERVATION AND CHANGE
CONDENSE
DISSOLVE SUBLIMATE BOIL
MELT
BREAK UP
SHAPE CHANGES SOLIDS LIQUIDS
DENSITY
OF DIFFERENT OBJECTS OF DIFFERENT MATERIALS
OF DIFFERENT OBJECTS SAME MATERIALS
OF AN OBJECT
PROPERTIES OF OBJECTS
----MINI - OBJECTS
GASES
MATTER OR NON-MATTER
PROPERTIES OF DIFFERENT OBJECTS
OBJECT - MATERIAL DIFFERENCES
MATERIAL - OBJECT DISTINCTION
MEASUREMENT AND DATA HANDLING
VOLUME-WEIGHT RELATIONSHIP GRAPHS AND RATIOS
CALCULATE AREA AND VOLUME
MEASURE WT, AREA AND VOL. DIRECTLY

Figure 13. A more detailed version of the relationships shown in Figure 7.

Some of the key practices that are enabled by scientific knowledge include the following:

- **Defining and describing**. Defining and describing involves recalling from memory a definition of a concept or principle or describing how one concept relates to other ideas. For example, a student could describe the flow of energy in an ecosystem. Or a student could describe how to use a light probe by telling a fellow student how to use it to measure light reaching a plant.
- **Representing** data and interpreting representations. Representing data involves using tables and graphs to organize and display information both qualitatively and quantitatively. Interpreting representations involves being able to use legends and other information to infer what something stands for or what a particular pattern means. For example, a student could construct a table to show the properties of different materials or a graph that relates changes in object volume to object weight. Conversely, a student could interpret a graph to infer which size object was the heaviest or a straight line with positive slope to mean there was proportionality between variables.
- **Identifying and classifying**. Both identifying and classifying involve applying category knowledge to particular exemplars. In identifying, students may consider only one exemplar (Is this particular object made of wax?) whereas in classifying students are organizing sets of exemplars. For example, they could sort items by whether they are matter or not matter; by whether they are solid, liquid, or gas; or by kind of substance.
- Measuring. Measuring is a simple form of mathematical modeling: comparing an item to a standard unit and analyzing a dimension as an iterative sum of units that cover the measurement space.
- **Ordering/comparing along a dimension**. Ordering involves going beyond simple categorization (e.g., heavy vs. light) to conceptualizing a continuous dimension. For example, students could sort samples according to weight, volume, temperature, hardness, or density.
- Quantifying. Quantifying involves being able to measure (quantify) important physical magnitudes such as volume, weight, density, and temperature using standard or nonstandard units. Predicting/inferring. Predicting/inferring involves using knowledge of a principle or relationship to make an inference about something that has not been directly observed. For example, students can use the principle of conservation of mass to predict what the mass of something should be after evaporation; or they may calculate the weight of an object from knowledge of its volume and the density of a material it is made of.
- **Posing questions**. Students identify and ask questions about phenomena that can be answered through scientific investigations. Young learners will often ask more descriptive questions, but as learners gain experiences and understanding they should ask more relational and cause and effect questions.
- **Designing and conducting investigations**. Designing an investigation includes identifying and specifying what variables need to be manipulated, measured, and controlled; constructing hypotheses that specify the relationship between variables; constructing/developing procedures that allow them to explore their hypotheses; and determining how often the data will be collected and what type of observations will be made. Conducting an investigation includes a range of activities—gathering the equipment, assembling the apparatus, making charts and tables, following through on procedures, and making qualitative or quantitative observations.
- **Constructing evidence-based explanations**. Constructing explanations involves using scientific theories, models, and principles along with evidence to build explanations of phenomena; it also entails ruling out alternative hypotheses.
- **Analyzing and interpreting data.** In analyzing and interpreting data, students make sense of data by answering the questions: “What do the data we collected mean?” “How do these data help me answer my question?” Interpreting and analyzing can include transforming the data by going from a data table to a graph, or by calculating another factor and finding patterns in the data.
- **Evaluating/reflecting/making an argument**. Evaluate data: Do these data support this claim? Are these data reliable? Evaluate measurement: Is the following an example of good or bad measurement? Evaluate a model: Could this model represent a liquid? Revise a model: Given a model for gas, how would one modify it to represent a solid? Compare and evaluate models: How well does a given model account for a phenomenon? Does this model “obey” the “axioms” of the theory?

Figure 14. Examples of what evidence of understanding in science might look like (From Smith et al 2004).