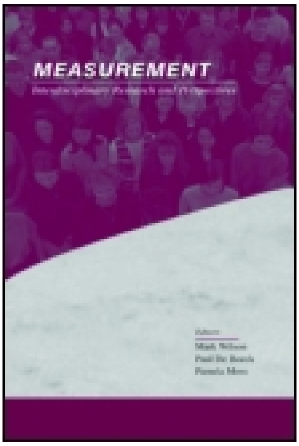


Measurement: Interdisciplinary Research & Perspective
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmes20

Road Maps for Learning: A Guide to the Navigation of Learning Progressions

Paul Black [a] , Mark Wilson [b] & Shih-Ying Yao [b]
[a] King's College
[b] University of California, Berkeley, USA

Available online: 29 Aug 2011

To cite this article: Paul Black, Mark Wilson & Shih-Ying Yao (2011): Road Maps for Learning: A Guide to the Navigation of Learning Progressions, Measurement: Interdisciplinary Research & Perspective, 9:2-3, 71-123

To link to this article: http://dx.doi.org/10.1080/15366367.2011.591654

Measurement, 9: 71–123, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2011.591654

Psychology Press
Taylor & Francis Group

FOCUS ARTICLE

Road Maps for Learning: A Guide to the Navigation of Learning Progressions

Paul Black
King's College

Mark Wilson
University of California, Berkeley

Shih-Ying Yao
University of California, Berkeley

The overall aim of this article is to analyze the relationships between the roles of assessment in pedagogy, the interactions between curriculum assessment and pedagogy, and the study of pupils' progression in learning. It is argued that well-grounded evidence of pupils' progressions in learning is crucial to the work of teachers, so that a method is needed which will enable the production of such evidence in relation to the learning strategy of any teacher. The argument starts by proposing a rationale for understanding the central roles of assessments in pedagogy and in particular the relationships between their use for formative and summative purposes. This is then related to a more general discussion of the links between curriculum, assessment and pedagogy which serves to highlight the importance of models of progression. The next step is to consider how assessment evidence of pupils' learning can be analyzed in two ways: By ordering the respondents in terms of overall scores, and by ordering individual items in terms of their difficulty. A method of relating these two in the BEAR assessment system is then explained. This method is then illustrated by a general review of the literature on the study of the atomic-molecular model, leading to detailed consideration of progression in the understanding of melting and evaporation. Results obtained, from a test of eleven items about these two topics, attempted by 665 grade 8 pupils in 11 schools in San Francisco, are then used to illustrate the method of analysis and the nature of the results that it can produce. A final

Correspondence should be addressed to Mark Wilson, Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720-1670. E-mail: MarkW@berkeley.edu

section considers the educational and assessment issues about learning progressions, pedagogy and assessment that we see as being informed by the ideas and practices outlined in the article.

Key words: curriculum, pedagogy, assessment, formative, summative, progression, Wright map, molecular model, changes of state

The concept of progression in pupils' learning has been a focus, implicitly or explicitly, of many research studies into pupils' learning, notably in science. The first aim of this article is to present a way of looking at problems inherent in classroom pedagogy, problems which have been made more severe by currently dominant forms of accountability systems. These current dominant forms have undermined what should be the normal and healthy relationship between formative and summative assessment, leading to a distortion of their natures and roles within the educational system. Specifically, under the regime of state accountability systems such as No Child Left Behind (NCLB) in the United States and the National Curriculum Assessments in the United Kingdom, summative assessment has overwhelmed formative assessment and, thus, taken over the central guiding role in determining classroom pedagogy (i.e., instructional practices). In this form, both formative and summative assessments take on roles that together narrow the curriculum, and inhibit good classroom pedagogy. We argue that there is an alternative way that the roles of formative and summative assessment in the education of students[1] can be brought together and enhanced to further students' educational progress. This requires that both formative and summative assessments be based on a common "road map" that can serve as a "backbone" for a learning progression, and that both have been built to be consistent and supportive of that road map. This concept of a road map is a generic idea, having many possible manifestations; we discuss one particular manifestation below.

The methods used to explore such learning and to analyze the data so produced have varied, both in the spectrum between the qualitative and the quantitative and in the spectrum between the evaluation of existing practice and the development and testing of innovative methods. Such work shows that the resulting schemes of progression can vary between cultures and can be changed by innovations in teaching. Given this variation, an overall aim of research on learning progressions might be to produce methods—with examples—to explore the particular learning progressions that emerge in any one context rather than to arrive at an ideal map of progression to which pedagogy should conform in all contexts. Thus, the second aim of this article is to explain how the BEAR Assessment System (BAS; Wilson, 2005) meets this aim. However, the use of any such analysis must be linked to consideration of assessment in the work of pedagogy as a whole. Thus this requirement is in accord with the first aim of this article.

Broadly speaking then, the overall aim of this article is to discuss three topics and to analyze the interrelationships between them. These three topics are (a) the roles of assessment in pedagogy, (b) the problems arising in the forms of interaction between curriculum assessment and pedagogy, and, given the relevance to these two of the study of pupils' progression in learning, (c) the application of the BEAR system in a way that makes a significant contribution to the study of such progressions. This overall aim is tackled in four main sections as follows.

We start in Section 1 by proposing a rationale for understanding the roles of assessments in pedagogy and in particular the relationships between their formative and their summative

[1] Throughout this article we have used the term "student" and "pupil" as if they were equlvalent, i.e., no significant difference is implied by the use of one rather than the other.

purposes. This is followed in Section 2 by a more general discussion of the links between curriculum, assessment, and pedagogy. In each of these sections we look ahead to the relevance of assumptions about learning progression that arise in the context of the issues they discuss. Section 3 then examines the first phase of analysis that is needed to implement the BEAR system and, thereby, serves as a prologue to the discussion in Section 4: It discusses, merely as an example, a particular topic in the science curriculum (the atomic molecular model), reviewing evidence about learning progression in this topic and the particular empirical results that have emerged from the testing of pupils. Section 4 explains the BEAR system and illustrates the characteristics of the results it can achieve by applying it to a modest collection of pupil performance data on tests of the topic described in Section 3. In our final section, Section 5, we discuss educational and assessment issues about learning progressions, pedagogy, and assessment that we see as being informed by the ideas and practices outlined in the article.

1. THE ROLES OF ASSESSMENT IN PEDAGOGY

In their accounts of formative assessment, Black and his colleagues (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Black & Wiliam, 1998) set out a variety of ways in which assessment activities are integrated into teaching and learning. However, these were not placed in the context of a broader model of pedagogy[2] and were only loosely linked to assumptions about the conditions required to enhance effective learning. One commentator on the 1998 paper challenged this restricted perspective:

> This [feedback] no longer seems to me, however, to be the central issue. It would seem more important to concentrate on the theoretical models of learning and its regulation and their implementation. These constitute the real systems of thought and action, in which feedback is only one element. (Perrenoud, 1998, p. 86)

Another limitation was that, in practice, many teachers had to carry out assessment for both formative and summative purposes, and to respond to external pressures of summative testing. However, most theoretical treatments of pedagogy say very little about the roles of assessment and its effects in classrooms and on classroom work and do not incorporate assessment into their schemas. Given that our overall aim is to show that the BEAR approach can help establish synergy between formative and summative applications of assessment, we have to propose a model for pedagogy in terms of which our subsequent discussion can be framed. This is the main purpose of this section.[3]

To achieve this aim, we first present our perspective on the meaning and practical applications of formative assessment. This will be followed by a presentation of the model of pedagogy that we are using, and in terms of which we can then say more about the links between formative and summative assessment in practice

[2] Different authors distinguish between pedagogy and instruction in different ways. In what follows we shall treat the two terms as synonymous with both denoting comprehensive overviews of the design and execution of classroom teaching and its assessment.

[3] Another problem is that in the understanding of some teachers, and of many publishers, formative assessment is equated to frequent testing and not seen to be at the core of the learning process—a problem that may arise because of a narrow understanding of the term *assessment* (see, for example, Klenowski, 2009). Our discussions here will show that this is a serious misunderstanding.

Our Perspective on Formative Assessment

It is first necessary to be precise about the meaning and practice of formative assessment. The definition adopted here is as follows:

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

This specification has three important implications. One is that feedback should follow a three-way path: from students to teacher so that the teacher can understand the students' level of understanding; from teacher to students, whereby the teacher responds to challenge or to extend the students' ideas; and from student to student, inasmuch as students can help and be helped by mutual dialogue. A second implication is that the definition includes feedback by students in assessing themselves and each other. A third is that feedback can be enacted both through oral and written exchanges, and over various time scales. These three implications will now be explored in turn.

First Implication—Oral Dialogue

Oral dialogue is a central component of classroom work. Insofar as the aim of such dialogue is to promote active participation by all members of a class in exploring and, thereby, developing their understanding in discussion with a teacher and with one another, the value of a question or activity proposed by a teacher rests mainly on its potential to provoke a range of responses rather than on its value in achieving a precise diagnosis of a particular step in understanding. The activity is then formative if the teacher can orchestrate and catalyze the variety of responses by encouraging students to clarify, compare, challenge and defend their various views to one another, so that they can begin to appreciate the strengths and the short-comings of their beliefs and can be led to modify them where necessary. The teacher plays a subtle variety of roles in this, being at appropriate points either the boundary setter for rules of argumentation, or a challenger, a provocateur, or a summarizer. Success overall then depends *first* on the power of the opening questions or activities to provoke rich discussion but, then, *secondly* on the capacity of the teacher to listen, to interpret the responses, and to steer the discussion with a light but firm touch, by summarizing, or by highlighting contradictions, or by asking additional questions. To do this skillfully and productively, one essential ingredient for a teacher is to have in mind an underlying scheme of *progression* in the topic; such a scheme will guide the ways in which students' contributions are summarized and highlighted in the teacher's interventions and the orientation the teacher may provide by further suggestions, summaries, questions, and other activities.

A dialogue that serves the learning of those involved must be characterized by reasoned argument, and a principal rule that the teacher must set is that assertions must be supported by arguments, with the corollary that someone else's argument can only be challenged in relation to weaknesses in the assumptions, reasoning, or evidence that support it, and in the validity of that reasoning in relation to the content under discussion. One cannot do this without first understanding these assumptions and inferences, which means that one must learn to listen carefully and thoughtfully rather than leap to rebut. Thus, careful listening is an essential feature of a learning dialogue, and cultivating such listening skills may be an important contribution to the learning of participants, even those who may say little at the time. However, such practice is not a characteristic of many classrooms (Applebee, Langer, Nystrand, & Gamoran, 2003; Smith, Hardman, Wall, & Mroz, 2004)

All of this is justified by a belief that discussion is an important component of learning and that the social aspects of learning are both essential and valuable. As Alexander (2006) expresses it:

> Children, we now know, need to talk, and to experience a rich diet of spoken language, in order to think and to learn. Reading, writing and number may be acknowledged curriculum "basics," but talk is arguably the true foundation of learning. (Alexander, 2006, p. 9)

He also draws attention to the contrast between countries where teachers believe that to be engaged in talk is intrinsic to anyone's learning, and those where they believe that it is marginal or preparatory to the "real" business of learning which is in writing.

Second Implication—Peer Group Dialogue

In most classrooms where peer group dialogue is common, the work alternates between whole class activity and students' discussion in small groups (see, e.g., Black et al. 2003; Mercer, Dawes, Wegerif, & Sams, 2004). Most of the argument above applies to such discussions, which are promoted in the belief that there is a distinctive value for learning in engaging with one's peers, using language and thinking that is at student level. There is also a clear advantage in that such a group can involve its members in more personal interaction than is possible in a whole class. What such activity can achieve in addition is the heightening of awareness within each individual of where their own views lie in relation to a spectrum of possibilities displayed by their peers. A group can be helped to achieve such awareness if they appraise and compare one another's written work (seatwork, homework, or test papers). Things can move further if such a group has to, for itself, place such pieces of work in some sort of (perhaps partial) order (of sophistication, say, or of validity in expressing understanding) and justify that ordering, for all are then involved in formulating and agreeing on criteria of quality. A group might then be helped further, with teacher guidance, to develop a clear view of the aim of a piece of work and of the ways in which criteria of quality can map out and so guide steps towards that aim. Such development contributes to students' metacognition, specifically to the power of overview of one's learning and, so, to the development of each student's learning autonomy.

Third Implication—Written Work

Much of the above is applicable also to consideration of the formative role of written work in supporting learning. Feedback in terms of comments which guide and require from the learner further work to improve on the work already produced makes full use, in the promotion of learning, of the time invested in the task. Yet here again, (a) to formulate a task or test so that the responses can provide evidence of learning progress, (b) to formulate helpful comments, tailored to the individual needs of each student, and (c) to give clear guidance on how to improve, all require a clear road map, that is, a view of the learning aim and of the steps along the route, or routes, that the student needs to take to get closer to that aim in light of his or her position en route. Furthermore, full student involvement requires that the students have some grasp, albeit perhaps imprecise, of the point they have reached along that route, for some students easily lose track and need to be helped by comments that remind them of what they have already achieved. The feedback must also give the student a clear aim for improvement, and if each student can locate this aim in a criterion-referenced framework, this can provide both orientation and motivation for improvement. With such an approach, each student will be competing against him- or herself, but may inevitably see him- or herself also in relation to peers and as competing with them.[4] These two foci of competition are highlighted in research into balance between the *task-involvement* and the *ego-involvement* of learners. The research studies of Butler (1988) and Butler and Neuman (1995) on written feedback, and of Dweck (2000) on self-theories, show both that the promotion of a culture of competition between students—through continual labeling of work with marks or grades and the publication of test or quiz scores—can shift students' attitudes in the direction of ego-involvement and that such a shift leads to poorer test results and to damage to their approach to learning. So whilst it is inevitable that some element of competition between students will always be present, a culture of collaboration rather than of competition, including attention to one's progress against criteria of progression, is to be encouraged. A similar message emerges in respect to peer-group work: A metaanalysis by Johnson, Johnson, and Stanne (2000) shows that collaborative groups produce significant learning gains, whereas where there is intra-group competition, group work produces no advantage over individual learning.

These considerations apply to peer group work, which forms part of the organization of classroom discussions as well as to work devoted to discussions of written work. In both contexts, it is clear that to participate in such work, students have to behave within some rules for productive group discussion: Mercer et al. (2004) emphasize, particularly, the importance of a rule that says that when making any assertion or counter-assertion a student must give a supporting reason. But it is also clear that the teacher's formulation of the written task or test, and of the guidance given about aims and criteria, has to be made in the light of the teacher's own clear view of the aims to be achieved for the topic and of the likely routes for progression towards those aims. In respect both of this strategic view, and of the conduct of a rational learning discourse, the way the teacher interacts with the learners will serve as a model for the way they should work in groups (Webb et al., 2006).

Formative Assessment: A Focal Learning Activity

There is strong research evidence that the various approaches to formative learning that are described above do improve students' attainments (Black & Wiliam, 1998; Wiliam, Harrison, & Black, 2004). One reason for this is that the principles of learning that underlie and are implemented in these activities are well established and firmly supported by cognitive research. These principles are, put simply:

- Start from a learner's existing understanding.
- Involve the learner actively in the learning process.

[4] A culture of competition is closely related to a norm-referenced perspective in assessment. In such a perspective, the overall performance of a group cannot improve, and the successful progress of some must lead to apparent regress of others whose achievement has not, objectively, changed. A criterion-referenced approach doesn't necessarily entail either of these consequences.

- Develop the learner's overview, i.e., metacognition—this requires that students have a view of purpose, an understanding of criteria of quality of achievement, and self-assess.
- Emphasize the social aspects of learning (i.e., learning through discussion) as these make a unique contribution.

The overall message of this section is that formative assessment must be understood as one of the fundamental core activities in the work of teachers. However, the next issue to consider is the broader context of pedagogy in which such assessment plays a role in order to discuss the characteristics of this context that are essential to the realization of its potential to enhance learning.

A Model for Pedagogy

The argument starts from an overview of pedagogy, focused on a sequence of the five steps that we see as being involved in the design and implementation of any learning exercise. Rather similar models have been proposed by Hallam and Ireson (1999) and Wiske (1999). Other authors have discussed some of the five steps, but not all of them: For example, Tyler's 1949 book discusses the first two and last two but says almost nothing about the third, the implementation step. The first three of these summarize and interrelate elements of the preceding discussion and the fourth follows naturally from them. But the fifth adds an extra dimension. In the applications of this model, it is envisaged that any account of the work of teachers will reveal interactions in both directions, that is, findings arising in the practice of any one step might lead to amendments to the work in a preceding step.

Strategic Aims

Here there can be wide, and legitimate, differences between different teachers, different schools, and different communities. For example, priority may be given either to understanding the concepts and methods of a particular subject, or to developing pupils' reasoning skills. For the former, a clear idea of an appropriate starting point and of the best route to take for developing this understanding are essential strategic requirements, whereas, for the latter the topic itself may only serve as a suitable context for furthering the development of particular reasoning skills. A lesson conducted to achieve the first of these aims would develop in a quite different way from one giving priority to the second. Of course, *both* content and reasoning skills may be emphasized, as well as the complex interactions between them. This issue is discussed in more detail in Black and Wiliam's 2009 paper.

Planning the Teaching

In this step, the tactics for realizing the strategy have to be chosen. As pointed out above, one important criterion here is the potential of any activity to elicit responses, whether oral or written, that help build a picture of learners' existing understanding, especially with respect to the learner's location on the learning progression, so that the next challenge can be framed to take that understanding further. Such design calls for foresight concerning the possibility of generating the interest and engagement of a class in relation to both oral dialogue, and to the interactions that might develop in relation to various forms of written work. The tasks so designed may range from questions that are narrowly diagnostic in their nature, to checking specific points, to tasks that are more broad in nature in order to explore more the full range of students' thinking.

Implementing the Teaching Plan

This aspect has been explored in the sections above on oral dialogue, peer group dialogue, and written work, which make clear that the way in which any plan is implemented in the classroom is crucial. The two problems here (discussed more fully in Black and Wiliam, 2009) are that for many teachers what is required is a re-appraisal of the way they perceive their roles as teachers and that the reactions of any group of learners are often unpredictable, so that adaptation of the teaching plan—in the light of the scheme of progression on which it is based—may be required.

Review—Informal Testing Serving Formative Purposes

At the end of any learning episode, there should be review, to check before moving on, perhaps using an end-of-topic test or other forms of assessment. Here there can be a dual purpose. One purpose is reflective, to both develop the learner's overview of the progress made and to check for gaps or misconceptions—overall, to serve as a progress review en route rather than as a terminal assessment. The other purpose is prospective, to look forward to building up a record of achievement, which might be a preparation for, and/or a contribution to, step 5. These purposes serve to distinguish this step from step 3, but such assessment can be used for formative purposes. Thus, what is revealed by the results of such tests may be a need for the whole class or for some particular students to repeat the work of step 3 to correct some faulty or incomplete understandings. Indeed, for this purpose, a review test will not be well used if time is not allowed to profit from its findings before moving on.

Questions used for this purpose may be open ended, requiring students to fully explain their responses, or they may be multiple-choice, where the distractors are crafted to express common responses so that students (a) can give an indication of their learning and (b) can have an opportunity to advance their own learning if they are expected to explain their reasons for accepting or rejecting each of them (see, e.g., Briggs, Alonzo, Schwab, & Wilson, 2006).

Summing Up—Formal Testing for Summative Purposes

Any test, or other form of assessment, is not in itself either formative or summative: The distinction lies in the way that the outcomes are interpreted and used. Thus, there may be overlap between steps 4 and 5 in the instruments used, but a distinction is to be made in the main purposes for which they are designed and their results are used to serve.

Summative assessments can serve several purposes. For individual students, and their parents, they can review progress overall and so can provide celebration, motivation, or guidance in relation to what has been achieved; they can also be passports to a further phase of education or employment. For the students' teachers they can provide an overall, perhaps externally calibrated, guide to the effectiveness of their work. For the school, they can guide decisions about the optimum choice of the next grouping and subject choices for the students involved, whilst for both a school and a public authority they can be used as a tool to inform the accountability of the teacher and/or school. It follows that some assessments, for example, those that lead to a certificate of completion for a student, might be only summative for the student, but might also be formative for the teacher. Indeed, the use of the results outside of the classroom can lead to high-stakes pressures on teachers and on students and can raise issues of pressure, trust, and resources. A common feature here is that results are used to inform decisions about future work.

It does not follow that a summative result has to sum up an extensive domain in terms of a single mark or grade; for at least some of the purposes they serve, summative results might well have to be detailed, for example, as a profile yielding separate data for each component of a set of domains of interest reflecting the stage of progression reached in each.

The central point of this article is to argue for a way of providing a strategic plan for developing a learning progression that will give a coherent framework to underpin these steps. Such a framework must provide the essential basis for steps 1 and 2, serve as a guide for the on-the-fly decisions that have to be taken in step 3, and provide the criteria on which the work in steps 4 and 5 should be based. For example, insofar as it has implications for further learning programs, a summative assessment ought to be framed in relation to a model, or road map, of progression in the learning domain, but if the purpose is to summarize over a more or less extensive period of learning, it may only need to reflect a model at the macrolevel, rather than at the microlevel, needed for formative work. If the same model of progression is used in steps 3, 4, and 5, a helpful relationship can be developed between all three. For example, the detailed guidance given to learners, and used by them, in the implementation (step 3) will be reflected in the design and scoring of the informal testing in step 4; then learners can be confident that the same framework will inform the formal tests of step 5 and the implications of the summative results for the teachers will similarly be in harmony with the design of their practice.

2. CURRICULUM, ASSESSMENT, AND PEDAGOGY

In order to develop our argument further, it is necessary to take a broader view of the problems discussed in Section 1 by saying more about step 1 of the scheme presented there. What are involved here are the overall aims of the learning in terms of the understandings and knowledge of the content that this learning is designed to achieve, that is, what is at issue is the curriculum. The importance of the interrelationships between curriculum, assessment and pedagogy (CAP) is an accepted feature in analyses of teaching and learning. Whilst the previous section brought into focus the interactive link between pedagogy and assessment, what has to be addressed here is that this interaction is strongly influenced by the nature of the curriculum and by the way it is interpreted within the overall context of schooling.

One possible, and common, scenario arises when both assessment and pedagogy have to work with a weak model of the curriculum. Insofar as a curriculum is little more than a catalogue of desirable outcomes (as it is with now-common “standards” approaches), it inhibits step 3 by giving no guidance in helping pedagogy to be designed in ways that foster student understanding, and it weakens step 4, thereby, inhibiting the design of assessments that can generate and support the interpretation of responses that lie between wholly correct and entirely wrong. Where assessment systems are designed to reflect a weak model of the curriculum without regard for the pedagogy, as will typically be the case in a high-stakes accountability environment, and if, in

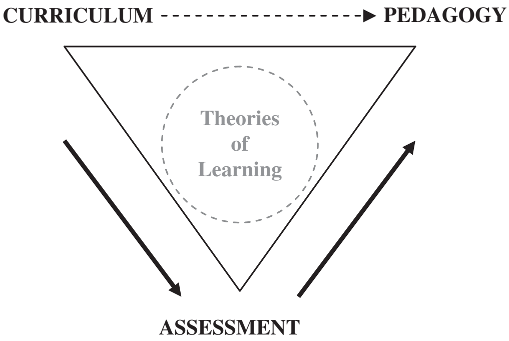


FIGURE 1 The breakdown of the standard interpretation of the learning triangle—the "vicious triangle."

addition, assessment tools are also weak, then the assessment that is produced will lack validity, in that it does not match to any model of learning more subtle than a dichotomy between "got it" and "not got it." Then, as Figure 1 shows, influence flows to pedagogy from the assessment, rather than vice versa, and teachers have to reconcile the pressures of external testing with their own ways of implementing the curriculum aims.

In effect, the power of interpreting a weakly specified curriculum rests with the testing agencies, and hence teachers will not be able to play any direct role in step 5. These testing agencies (and the policy makers who determine budgets) will naturally tend to prioritize the properties of efficiency and reliability in the systems and instruments they use; the agencies will have very weak links with teaching and learning issues and will not experience at first hand the effects on teachers, on pupils, and on classroom learning that their practices will produce. At the same time, teachers may well feel that they have to "teach to the test," which will be to the detriment of good learning practices. So steps 3 and 5 will be in tension, and practices in step 4 can either mirror those of step 5, so weakening the work of step 3, or remain faithful to step 3 and serve as poor preparation for step 5. There will be a very strong temptation to simply copy or mimic the external tests for their own summative assessments, and, hence, for such people there will be little motivation to develop their own summative assessment skills.

The consequences for pedagogy are then twofold. For their day-to-day work in schools, teachers can only find or fashion weak tools to explore or extend the imperfect responses of their students. At the same time external summative tests will not be designed to usefully reflect intermediate stages in the progress of pupils towards full understanding, so that they will be of little or no value in guiding the day-to-day work of classroom learning. Under accountability pressures, the curriculum comes to be defined in practical terms by the summative test, and the teacher is driven to ensure success by teaching to the test, since teaching for understanding may yield no reward if the test is not sensitive to that understanding (and all the pressures of cost and time efficiency militate against the possibility that the summative tests are indeed sensitive to understanding).

Thus the interactive formative assessment of step 3 is constrained both by the lack of suitable tools and by the prospect that its implementation may be judged to promise few rewards. The overall consequence is that the pedagogy is driven by assessment and that the curriculum exerts its influence mainly through its effect on assessment. This is indeed the major thrust of the implementation procedures for the No Child Left Behind legislation in the United States and the National Curriculum Assessments in the United Kingdom.

For such a model, as represented in Figure 1, any theory of learning can have only a weak effect on the translation of the curriculum into pedagogy but a stronger effect on the link between assessment and pedagogy, with assessment being used mainly as a means to diagnose the understanding of the students as individual learners. In this approach, the interaction of theories of an individual's cognition with complex psychometric analysis of responses may strengthen the diagnostic element. But without a direct connection to a rich pedagogy, such feedback will usually be inadequate to guide the day-to-day business of pedagogy.

This "vicious triangle" can be replaced by a better approach. A first requirement, as we have emphasized above, is that the curriculum needs to be fashioned in terms of a model, grounded in evidence, of the paths through which learning typically proceeds as it aims for the desired targets, that is to say, the curriculum reflects and provides a strong model of progression in learning. This "road map" may then inform both pedagogy and the assessments of steps 4 and 5, in that an articulated set of tools can be tailored to stages in progression along the road, so that such tools will help to identify the region along the road where failure gives way to success. The assessment is seen as more valid in that it is an indicator of a student's progression in learning. Thus the first issue to be addressed is to reform the interaction between curriculum and assessment, a reform that should be strongly driven by theories of student learning as in Figure 2, but also strongly influenced by the observation and interpretation of student growth as represented in the analysis of student responses to assessments.

The second issue is to develop the use of the road maps, formulated from the explorations through assessment and the interpretation of students' responses, as guides to the pedagogy. They will serve this purpose well if teachers are closely involved in the final summative judgments of their pupils. If teachers share responsibility for some of the tasks on which this final judgment is based, they can both match these to their own formative work and also influence the formulation of these tasks, and the criteria by which they are assessed, so that the results are valid in reflecting and motivating good learning practices. However, this can only be achieved if teachers are given the training in the formulation of such tasks that many will require. Black,

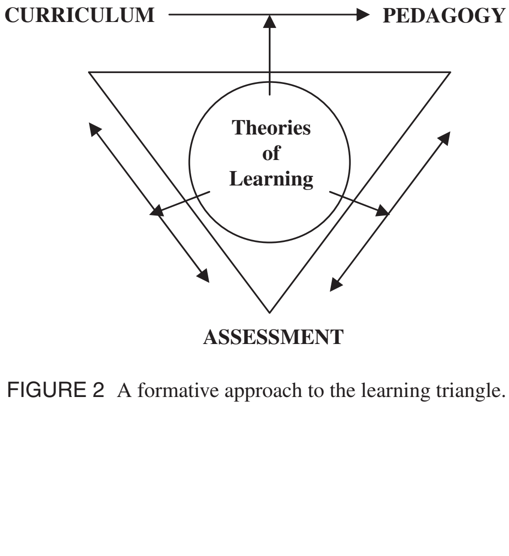


FIGURE 2 A formative approach to the learning triangle.

Harrison, Hodgen, Marshall, and Serret (2010, in press) give an account of work developed for this purpose. Teachers will also be well placed to explain the assessment regime to their pupils and to show that there is continuity between testing events and normal classroom learning. For example, as pupils work on an open-ended project task, teachers can ensure that it is sufficiently similar to work done previously so that they can be confident that every pupil understands what is expected. In addition, as pupils proceed with a task, a teacher can give help to any pupils who encounter obstacles that might prevent progress and that might, thus, mean that that these pupils will not give a full account of their capabilities; at the same time, the teacher will be in a position to allow for that extra help in making the final overall judgments.

For the many uses of summative results, for example, certification of pupils, public interest would require that an outside agency oversee such work by, both setting rules and guidelines for the nature of teacher-assessed tasks and for ensuring comparability between the judgments of different teachers and by, perhaps, supplementing the data from these with results of externally set and marked tests. Such an agency would have to work closely with teachers in order to take into account both the interpretations by teachers of the aims and content of the curriculum and the teachers' experiences of working towards these aims with their pupils. There are state systems, notably in Australia, where this has been done (Stanley, MacCann, Gardner, Reynolds, & Wild, 2009). Overall, the work done in step 4 will face both ways, back to step 3 for formative implementation of the lessons learned in informal reviews and forward to step 5 in that there will be continuity between this formative work and the tasks and criteria that will be highlighted for step 5.

The road maps will underpin all of the steps. As guides they will not only help strategize the planning, whether on the macroscale of step 1 or the microscale of step 2, but they will also help the actual execution in step 3 in that the flexibility required by a formative approach will draw upon the road maps to optimize the steering along the "road." In this formative pedagogy, theories of learning will have a crucially strong influence, although what will then be drawn upon will include theories of learning through discussion of group collaboration, of developing metacognition, and of the value of enhancing task-involvement rather than ego-involvement.

Note that in developing a set of learning practices (e.g., combinations of curriculum, pedagogic, and assessment practices) based on the formative learning triangle shown in Figure 2, the developers will need to engage in wide-ranging studies of the success of those practices. In these studies, the results of the pedagogy will reflect back, via the assessment, to the curriculum, and, thus, the arrowheads will point both ways from pedagogy to curriculum and back via assessment.

The synergy represented in Figure 2 does not of course resolve all assessment problems. Reliability is not addressed. This is not the same problem for formative work as for summative work, because in any given step in progression, far more evidence is collected than could be elicited in a summative exercise, and the consequences of wrong interpretations can be quickly manifest so that the mistake can be put right. However, it is a serious problem for summative assessment. The fact that a coherent model interrelates the assembly of assessment instruments may ameliorate the problem (Stanley et al., 2009; Black et al., in press)

For validity, tools that faithfully reflect the characteristics of the "road map" will likely enhance the validity of summative judgments, but only if the nature and range of these tools is adequate. Thus the (relatively) short formal test employing only multiple-choice questions might be better constructed but might still be inadequate. In particular, where a learner may have made progress to the stage of a useful but flawed understanding of an idea, it will be far easier to see and appraise this in a constructed-response question than in (say) a multiple-choice item, because the learner's explanation will provide information about the nature of any misunderstanding.

3. A ROAD MAP FOR THE ATOMIC-MOLECULAR MODEL

As stated in the introduction, one purpose of this section is to serve as a prologue to Section 4: It discusses, merely as an example, a particular topic in the science curriculum, reviewing evidence about learning progression in this topic and the particular empirical results that have emerged from the testing of pupils. It will thereby illustrate the procedure for preparation of a road map by discussing the results of work with pupils that provides a basis for developing their understanding of the atomic-molecular model of matter. This presentation will focus on one component of a larger overall map that aims to present an over-view of progress towards the understanding of this model. This overall map is shown, in outline only, in Figure 3.[5] The full map would include far more detail within each of the component boxes. The overall purpose that is envisaged is to lay a foundation of evidence about matter so that a hypothesis matter is composed of large numbers of tiny identical particles might seem reasonable and, in particular, might not be seen to contradict the everyday evidence of their senses.

The particular component considered to serve the purposes of illustration in Section 4, is box 4 of this map. Work in this box would help lay the foundation for work on the atomic-molecular theory in boxes 5 and 6. It is assumed here that the atomic-molecular model is introduced as a hypothesis only at the end of the teaching sequence. Since all students will have heard of atoms

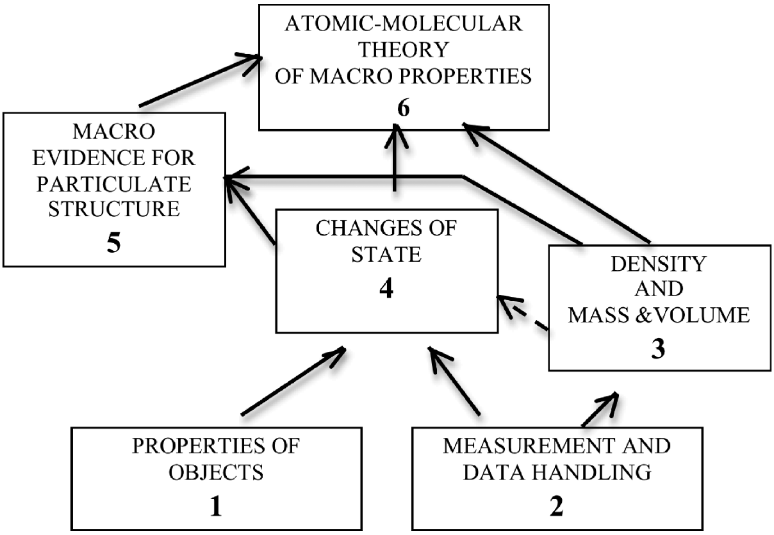

ATOMIC-MOLECULAR THEORY OF MACRO PROPERTIES
6
MACRO EVIDENCE FOR PARTICULATE STRUCTURE
5
CHANGES OF STATE
4
DENSITY AND MASS &VOLUME
3
PROPERTIES OF OBJECTS
1
MEASUREMENT AND DATA HANDLING
2

FIGURE 3 Outline of road map for molecular theory of matter.

[5] In this diagram, and throughout this article, progression will be envisioned as proceeding from the bottom to the top of any relevant diagram or table.

and molecules, this may seem unrealistic. Indeed, there is disagreement in the literature between those who argue that teaching had best present the atomic-molecular model at an early stage (the "molecules-first" approach recommended by Papagiorgiou & Johnson (2005) and Johnson & Papagiorgiou (2010)) and those who argue that early work should focus on the evidence that will later serve as the basis for justifying a particle model (the "particles-first" approach of Nakhleh et al. (2005) and Prain, Tytler, & Peterson (2009)). The point is however that there is ample evidence, as we explain below, from the science education literature on misconceptions, that pupils may not be able to reconcile this pre-existing fragmentary knowledge with other things that they know about matter. To correct this, it is necessary to build a basis of evidence that will interlink the fragments in a coherent way so that the overall picture can be seen to support the main hypothesis in a persuasive way and without raising contradictions.

The substructure within each of the six components represented in Figure 3 will not be discussed here; for our present purpose it is only necessary to show the detail of box 4. The purpose of the learning of the topics in this box is to establish that there is continuity of the substance involved over a range of changes in the macroscopic, observable properties of samples of a substance. Our choice here does not imply that topics that belong in the other five boxes are irrelevant—indeed the ideas explained in boxes 1, 2, and 3 would underpin work in box 4. A starting point for the box 4 topic, as treated in box 1 (of Figure 3), should be that materials can be broken up into smaller and smaller bits and still retain many of their properties, so there is dependence on this component, the function of which is to present experiences and develop thinking about this issue. As changes of state are explored, it is necessary to consider, at an early stage, changes in mass and volume (box 3), so there is also dependence on component 2, Measurement and Data Handling, which serves to develop acquaintance with the collection and analysis of measurements of mass, length, and volume. There is a further dependence on some elements of component 3, the ultimate function of which is to develop the concept of density as a common property of different pieces of the same material (a concept that will be needed in components 4, 5, and 6). The relative orientation of the boxes for numbers 1, 2, 3, and 4 in Figure 3 gives a rough indication of these levels of dependence; a full diagram might well indicate more inter-connecting links. For example, evidence about changes of state (box 4) provides only one source of support for the idea that matter is composed of large numbers of tiny identical particles; experiences discussed in other boxes may also be directly relevant. Some of the sources listed in Appendix A consider topics in box 4 together with linked topics in boxes 1, 2, and 3.

Likewise, other researchers have studied links with and the contents of boxes 4 and 5 (see, e.g., Shawn, Delgado, & Krajcik, 2010). For box 5, Macro Evidence for Particulate Structure, evidence about changes of state, in cases in which one considers only one material at a time, should be supplemented by evidence, notable from the mixing of liquids and from the phenomena of dissolving, to build as full a picture as possible about the relevant basis, in macroevidence, for work to start on component 6. Here again, this will draw strongly on component 5, but will also develop the arguments by drawing on ideas established in several other components.

For a map at the level of detail we consider for box 4, there is no implication that the relative importance of, or the teaching/learning time needed for, the different components is represented by the relative sizes of the boxes in the diagram. Also, no particular sequence *within* each box is implied; it may or may not help the learning to complete the whole of one component in a single unbroken sequence. It is relevant here that the interlinkages between any one component and another may be multiple, and that any one may contribute to more than one of the others.

This makes clear that the progress maps, seen in isolation from one another, are building blocks for a teaching scheme, not in themselves a proposed teaching sequence. However, any scheme has to introduce the prior-dependent links between one construct and another.

The first stage in the development of a map such as that presented in Figure 3 (and the more detailed maps in Appendix B of the internal structure of any one component that will be shown later) will be based upon the synthesis of several sources of evidence. These would include the following:

- research results about pupils' common misconceptions
- the internal logic of the concepts involved—any one idea may depend on another
- indications from learning theory about the difficulty of the types of thinking involved
- results from assessment items that indicate problems and possibilities with the topic sequence

These four points will be discussed in more detail in relation to our detailed consideration of box 4, Changes of State.

Research Results about Pupils' Common Misconceptions: Overview[6]

An early comprehensive review (Andersson, 1990) pointed out that a clear concept of "state" was not grasped by young children. Gas, for example, was exhaust gas or war gas; air is not gas and has no weight. Particles might be seen as small pieces of matter and the change from macro- to microparticles no more than a process of subdivision— whilst atoms are not matter but are *in* matter, like raisins in a bun. The review was critical of textbooks, which both made inaccurate statements (e.g., an atom is the smallest bit of a substance) or discussed a concept (e.g., conservation) as if it were self-evident.

That summary served to illustrate the range of issues involved in this field. The topic includes the solid-liquid transition, so including melting and freezing, the solid-gas transition (i.e., sublimation), and the liquid-vapor transition, so including evaporation, condensation, and boiling. Underlying pupils' ideas about these are their ideas about the three states of matter, about conservation and reversibility across transitions, about such concepts as substance, particles, and molecules, and about the ontologies and epistemologies involved, all linked or developed within their reasoning skills. Over the last 20 years, more than 25 papers have reported research on these topics, each study focusing on a particular subset of them. Most of these studies are cross-sectional across several ages; very few are longitudinal. A few involve more than 100 samples or less than 20—the majority being in the 20 to 100 range. About half involve some form of intervention in the teaching, but only a few undertake pre-post comparisons. The majority rely mainly on a qualitative approach using category analysis of interviews with pupils, the data analysis involving counts of numbers and of changes, with instruction or with age, in these. In cases of written responses being used, Bar and Travis (1991)and Bar and Galili (1994) reported that responses to open-ended questions could differ significantly from those for multiple-choice questions. There is little or no use of correlations or of analyses of variance. Table A1, in Appendix A, gives a summary of the research designs and methods used in the papers referenced here.

[6] The research studies used in this section are summarized in Appendix A.

In what follows, we first review two main themes—melting and freezing—and then evaporation, boiling, and condensation. We then describe the various dimensions along which progression schemes have been proposed in the literature. Finally, we look at a key disagreement between some who argue that particle ideas are essential to understanding the phenomena and others who take the opposite view.

Research Results about Pupils' Common Misconceptions: Freezing and Melting

Both Stavy (1990) and Ross and Law (2003) have shown that both middle school pupils and high school pupils see weight as a function of either the state of matter or its hardness—so ice weighs more than the water that froze, whilst chocolate decreases in weight when it melts because it's softer. Stavy argued that conservation of weight must be taught first because ideas of particle theory cannot be established on the basis of faulty ideas about such conservation, whilst Ross and Law argued that conservation of atoms could be a basis for understanding. Nakhleh et al. (2005) found that some middle school children talked of particles falling off a block of ice as it melted whilst others said that the ice spreads out into a pool of water; however, this work focused on the association of these descriptions with heat and with temperature changes, as did the work of Paik et al. (2004). The latter reported also that pupils over the age range 5 to 13 did not mention a particle model in relation to melting and freezing.

Research Results about Pupils' Common Misconceptions: Evaporation, Boiling, and Condensation

There are more papers on evaporation, boiling, and condensation than on melting and freezing. Several studies (Bar & Travis, 1991; Johnson, 1998a, b; Tytler, 2000; Tytler & Peterson, 2005) showed that pupils have difficulty with evaporation because they think that water cannot just disappear, and if air/gas is not seen as a substance, nothing can be up above the liquid surface. It is common for pupils to think that a liquid disappears into the substrate, or that it has been "carried up" into the clouds or into the sun; ideas learned about the water cycle support such accounts. Thus, there is no concept of water vapor and it is not recognized that steam is composed of liquid droplets. For boiling, pupils find difficulty in suggesting what is inside the bubbles within boiling water. Some say that they are bubbles of air; some, that water turns into air in the bubbles. Chang (1999) found that almost half of the students in a teachers' college course shared these various misunderstandings and were not clear about conservation of weight in changes of state; moreover, such weaknesses were common even amongst science specialists. Given all of this, it is not surprising that these same studies showed that condensation is found equally, or more, difficult. The importance of the context in such research explorations was highlighted by Costu and Ayas (2005), who found that for pupils across the high school age range, whilst 40% understood evaporation in an open system, only 17% understood it in a closed system,—many saw evaporation as a chemical change—and the responses differed for different liquids. An aspect highlighted by Varelas et al. (2006) was that pupils were confused by the apparent contradictions between everyday experience and their classroom work and by linguistic shifts between these contexts, so that they challenged the taught ideas, seeing them, for quite sophisticated reasons, as problematic. Paik et al. (2004) argued more fundamentally that pupils had to develop higher reasoning skills before they could even understand the concept of a "state of matter," whilst Johnson and Papageorgiou (2010) argued that a firm concept of a "substance" had to be developed first to replace the ambiguity of the term "matter"—the latter often denoting a mixture of substances.

Research Results about Pupils' Common Misconceptions: Dimensions of Progression

Several authors propose schemes for the progression in the learning involved for this topic. However, of those referenced here, only two are based on a longitudinal study and only seven are based on cross-section studies over several ages. Insofar as these schemes cover different dimensions of progression, they cannot be compared and any attempt to align them would seem problematic.

The schemes of both Nakhleh et al. (2005) and Johnson and Papageorgiou (2010) concern understanding of particles: their sequences propose that pupils develop from having no sense of particles (a macrocontinuous model of matter), to seeing particles as embedded in matter, to equating the properties of the particles to the macroscopic properties on a substance, and, finally, to understanding that the properties of the microscopic particles are not the same as the macroscopic properties.

The schemes of Andersson (1990) and Bar and Galili (1994) set out broadly similar suggestions for progression in the understanding of evaporation. They focus on the explanations, proposing the following sequence of pupils' accounts of what happens to the water: disappears; is absorbed; is dispersed; is transferred (to clouds, etc.); is modified or undergoes transformation.

Stavy's (1990) sequence arose from her work on ideas about weight; her sequence for development here distinguished gas from liquid. For gas, in relation to the liquid or solid it came from, the sequence is the following: it has no weight; weighs less than; weighs the same as (the liquid or solid). However, for a liquid, in relation to the corresponding solid, it weighs more than the solid; zero (liquids don't feel heavy); less than; equal to.

A fourth category focuses on development in the patterns of reasoning that are deployed. These are more diverse and more difficult to summarize. Stavy linked some of the changes observed to changes, which varied according to context, from declarative accounts to operative accounts. Tytler and Peterson (2005) argued that progression depended on changes across three dimensions: the conceptual/ontological, the epistemological, and the episodic. Their evidence was that from age 5 to age 11, pupils' arguments and perceptions became more abstract and more consistent; developed distinctions between matter and its properties; showed more accurate observation; and improved in cogency and consistency in the quality of reasoning. On this basis Tytler and Peterson proposed a scheme for evaporation similar to those of Bar and Galili (1994); it was shown in both of these studies that the patterns of change were complex and variable; that the reasoning of any one pupil could change from one context to another and could differ between open-ended question responses and choices in a multiple-choice item; that no single model of the changes could be adequate; and that explanations of any one individual were closely bound up with issues of personal identity. The difficulties pupils have in resolving differences between everyday experience and school explanations are also relevant here: Nakhleh et al. (2005) pointed out that secondary pupils were more varied in their uses of microlevel or macrolevel explanations than primary pupils, perhaps because of the variability of their development in the capacity to make the distinctions involved.

In the Piagetian perspective proposed by Shayer and Adey (1981), many of the issues involved in the present set of topics would call for their early-formal to their late-formal categories, so that it would be expected that full understanding would be achieved only in the upper middle school and high school stages and that performance would be diverse across each age group. The social-constructivist approach underlying our explanation of formative assessment should not be seen as inconsistent with this (see Shayer, 2003).

Two general conclusions emerge from this survey. The quality of the evidence in these studies is not strong, and the qualitative bias in the methods means that rich pictures emerge, but it is difficult to generalize from these. A more serious difficulty is that it might be expected that any findings might be dependent both on the sociocultural differences between different pupil samples and on the teaching content and styles these pupils had experienced; several of the studies say very little about these features. Only about one-third of the studies were intervention studies, in which the evaluation had been linked to a specific teaching scheme with well-defined aims. Finally, any evidence about progression seems speculative, because of the lack of longitudinal studies, so that they are based on overall average changes of groups rather than on trajectories of change of individuals.

Composing a Road Map for Changes of State: Mapping the Components a Priori

As noted above, the purpose of a learning program that includes work on changes of state is to establish that there is continuity of the substance involved over a range of changes in the macroscopic, observable properties of samples of a substance. The map we set out here draws on the literature summarized above, and on the review of Smith et al. (2006), including from the latter our analysis of some of the examples they quote.

For our purpose, we propose that understanding of the topic may be surveyed in the light of a general construct about the persistence of a material over gross (but physical not chemical) changes, notably, for this component, changes of state. There were three steps involved in formulating our hypothesized structure (see Figure 4).

The **first** step explores changes in appearance only. It includes change of shape of a solid or of a liquid by the breaking up of an object and change of shape by thermal expansion. The **second** step looks at changes of state, focusing only on the one material in a closed, or effectively

SYNTHESIS
S1. Reversibility and continuity

CHANGES OF STATE
CS3. Condensation
CS2. Evaporation, Boiling and Sublimation
CS1. Melting

CONSERVATION IN CHANGES OF APPEARANCE
CA3. Thermal expansion
CA2. Breaking up
CA1. Changes in shape

FIGURE 4 The learning progression in changes of state.

closed, system. It includes changes associated with melting, with evaporation and boiling, with sublimation, and with condensation. The **third** step aims to emphasize the theme of continuity of material across all the changes explored, an idea that is strengthened by evidence about the reversibility of the changes.

These three steps are set out in a sequence that assumes that understanding will best be developed by using the sequence presented here. It must be emphasized that whilst this scheme has emerged from our judgment of the best strategy that should be adopted as a framework for a teaching scheme, it is set out here as an exemplary case to which the analyses we describe might be applied. Others may propose a different scheme for the same topics: the analysis in our next section could equally well be applied to an alternative road map, a point discussed more fully in Section 5. However, we would stress that the approach, involving four principles linked to the four building blocks, as set out below, should be followed in developing any map.

The discussion in this section is also relevant to Section 4 in a different way. The review of the literature summarized in Appendix A has helped in the task of proposing a road map for further empirical exploration, but because of the diversity of assumptions, methodologies, and pupil samples, it could not do more than serve as a loose guide to that proposal. The argument between “molecules first” and “particles first” is one example of this diversity. Hence one question to be addressed is how the integrity of such a proposal, and ways in which it may be refined, can be evaluated by further empirical analysis of the outcomes of pupils’ learning. The next section describes a way to tackle this question.

4. THE BEAR ASSESSMENT SYSTEM

In this section we give an account of the measurement approach we will apply to the atomic-molecular learning progression, the BEAR Assessment System (BAS), and discuss each of the four principles and building blocks on which it is based. Throughout these descriptions, we will refer to examples from the atomic-molecular model.

The Berkeley Evaluation and Assessment Research (BEAR) Center has for the last several years been involved in the development of the BAS. The system is based on four principles, each associated with a practical “building block” (Wilson, 2005), as well as an integrative activity that can take on different aspects under different circumstances (e.g., assessment moderation/cut score setting). Its original deployment was as a curriculum-embedded system in middle school science (Wilson & Sloane, 2000), but it has clear and logical extensions to other contexts such as in higher education (Wilson & Scalise, 2006), in large-scale assessment (Wilson & Draney, 2005), and across other disciplinary areas such as chemistry (Claesgens, Scalise, Draney, Wilson, & Stacy, 2002). The BEAR Assessment System is uniquely suited to meeting the needs of the vision of this article because it is based on an overarching interrelationship among different levels of assessment (Wilson, 2004a; Wilson, 2004b; Wilson & Draney, 2004).

As implied above, all assessment systems are based, implicitly or explicitly, on models of student learning. Good models 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 relate observations back to learning models and ideas of cognition. There follows a second consideration: that to serve as a sound basis for evidence, the tasks themselves need to be systematically developed with both the learning model and subsequent inferences in mind and they need to be tried out and the results of the trials systematically examined with those same things in mind. There is a third consideration, namely, that these inferences should provide 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 learning model or the items should be (see NRC, 2001, for a full exposition of these principles).

The BEAR Assessment System addresses these considerations through four principles:

I. A developmental perspective
II. A match between instruction and assessment
III. Management by teachers
IV. Generating quality evidence

The overall aim is to help in the exchange of appropriate feedback, both from students to teachers and from teachers back to students, in the short term, and into their planning in the long term. The account given here of these four principles is laid out in more detail in Wilson (2005). The overall scheme is illustrated in Figure 5: it shows that each of the four principles is linked to a building block that spells out the practical implementation of the principle. Below we take up the principles in turn, explaining each and then describing the matching building block that implements the principle. For each building block, we highlight the implications for both formative and summative assessments.

Principle I: 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 criteria for a sound developmental perspective has been a challenging goal for educators for many years. From Bruner's nine tenets of hermeneutic learning (Bruner, 1996) to considerations of empirical, constructivist, and sociocultural schools of thought (Olson & Torrance, 1996) to

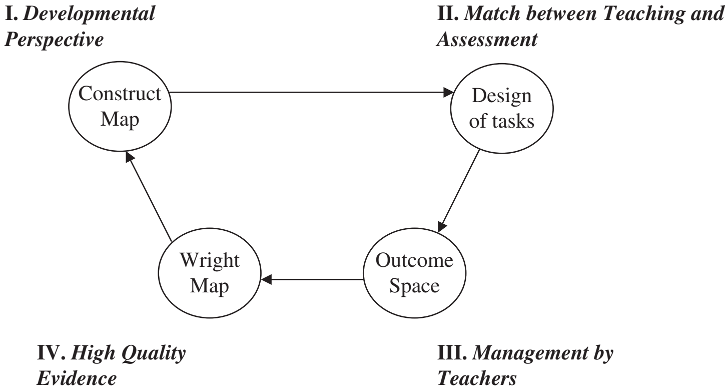

I. Developmental Perspective
II. Match between Teaching and Assessment
Construct Map
Design of tasks
Wright Map
Outcome Space
IV. High Quality Evidence
III. Management by Teachers

FIGURE 5 The *Principles* and Building Blocks of the BEAR Assessment System.

the very influential NRC report, *How People Learn* (NRC, 1999), 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 (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 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 both from the expertise and from the learning theories of the teachers and/or curriculum developers involved in the developmental process.

Building Block One: The Construct Map[7]

In the phrase *construct map*,[8] the term *construct* refers to the understanding of a concept that forms a significant and necessary step in the learning of the subject (similar to Meyer & Land's (2003) notion of a threshold concept). A construct map must be based on a coherent and substantive definition for the content of the construct. In addition, the construct must have a particularly simple form, in that it extends from one extreme to another, from high or strong, to low or weak, degrees of understanding of the concept. However, in the practice of assessment, this ordering can be looked at in two ways: what one is interested in finding is where the respondent is located in this continuum, but what one can observe is the responses of learners to the tasks they tackle. For example, inadequate responses of any individual may align with those of the group as a whole, or they may indicate an idiosyncratic pattern. The diagnosis of an individual's problem will differ between these two cases.

These different aspects of the construct, the respondents, and their responses lead to two different sorts of construct maps:

i. A respondent construct map in which the respondents are ordered from more-sophisticated to less-sophisticated understanding—and that qualitatively may be grouped into an ordered series of levels
ii. A task response construct map in which the item responses are ordered as evidence of more-sophisticated to less-sophisticated understanding—and that qualitatively may be grouped into an ordered series of levels.

The best type of construct map will, however, include both of these. Note that here we are using the term construct map for just one strand of what may be several strands that are needed to fully describe a learning progression.

[7] Note that the term *progress variable* is also used in the literature (e.g., Masters, Adams & Wilson, 1990; Masters & Forster, 1996)—the distinction between this term and *construct map* is that *progress variable* would be used to refer to the result of applying all four of the building blocks.

[8] The term *construct map* is the equivalent in the BAS of the "road map" that we have described above. Effectively, a construct map is one particular type of a road map (the one that belongs with the BAS). Hence, we will use *construct map* in the text regarding the BAS, and return to using the generic term *road map* in the section Discussion and Conclusion.

A generic construct map is shown in Figure 6—the construct being measured is called “X.” This idea will be used throughout this article, so we shall describe its parts before moving on. The arrow running up and down the middle of the map indicates the continuum of the construct, running from “low” to “high.” The left-hand side will indicate qualitatively distinct groups of respondents, ranging from those with low performance overall on X to those with high performance. A respondent construct map would include only the left side. The right-hand side will indicate qualitative differences in responses to specific tasks, ranging from responses that indicate low X to those that indicate high X. A task response construct map would include only the right side. A full construct map would include both. For any one task, there may be many different responses. Some such responses may indicate an equivalent level of understanding, that is, the same level of X, but some tasks may evoke responses that differ in that they are evidence of different levels of X.

As represented in Figure 6, a respondent with low X is likely to give a task response that indicates the lowest level of X and is less likely to give one that indicates a higher level; likewise, a respondent with high X is likely to give the response that indicates the highest level of X and is less likely to give responses linked to lower levels. As explained below, the link between the two sides will depend on the probabilities of responses.

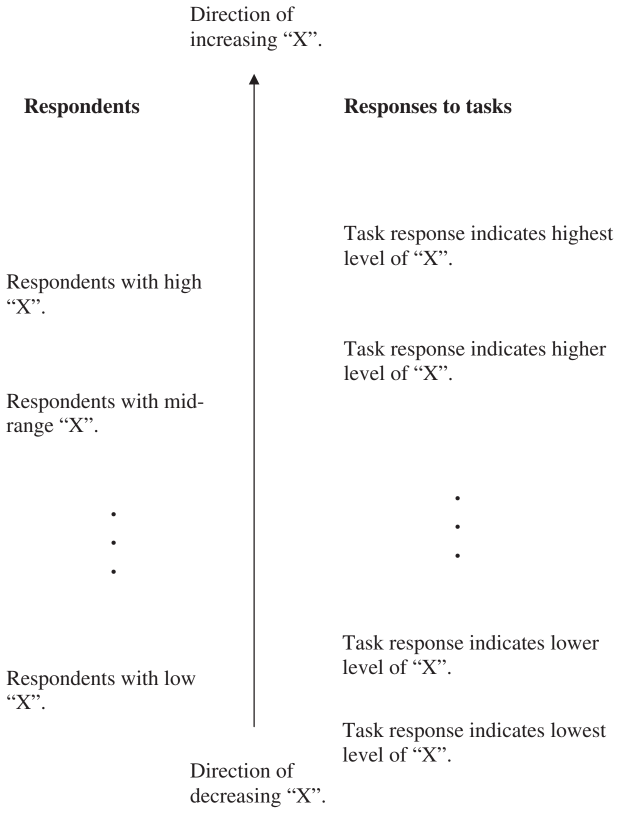

Direction of
increasing “X”.
Respondents
Responses to tasks
Task response indicates highest level of “X”.
Respondents with high “X”.
Task response indicates higher level of “X”.
Respondents with mid-range “X”.
Task response indicates lower level of “X”.
Respondents with low “X”.
Task response indicates lowest level of “X”.
Direction of
decreasing “X”.

FIGURE 6 A generic construct map of construct “X.”

Note that this depicts an idea rather than being a technical representation. Indeed, later in this section, this idea will be related to a specific technical representation, called a Wright map, but for now, we just concentrate on the idea. Certain features of the construct map are worth pointing out:

1. There is no limit on the number of locations on the continuum that could be filled by a student respondent on the left, or a task response label on the right. Of course, one might expect that there will be limitations in the accuracy of such location placements caused by limitations of data, but that is another matter.
2. The task labels relate to responses to tasks, not to the tasks themselves. Although one might tend to reify the tasks as phenomena in their own right, it is important to keep in mind that the locations of the labels are not the locations of tasks *per se* but are really the locations of certain types of responses to the tasks. The tasks' locations are represented via the respondents' reactions to them (thus making the two sides interdependent in their definitions).

There are several ways in which the idea of a construct map can exist in the more complex reality of usage; a construct is always an ideal—we use it because it suits our theoretical approach. If the theoretical approach is inconsistent with the idea of mapping in this way then it is hardly sensible to want to use a construct map as the fundamental approach; an example would be a case in which the theory was based on an unordered set of latent classes. But there are also constructs that are more complex than the construct map yet contain construct maps as components. Probably the most common would be a multidimensional construct. In this sort of situation, in order to use the construct mapping approach, it is necessary to focus on one dimension at a time.

Examples of construct maps abound: In educational testing, there is inevitably an underlying idea of increasing correctness, increasing sophistication, increasing excellence, and so on. We will illustrate the idea of a construct map by using the example that was described above: Changes in State (see Figure 4). In this learning progression, we postulate that students' understanding of changes in states of matter can be separated into at least three dimensions: (1) understanding of melting, (2) understanding of evaporation, boiling, and sublimation, and (3) understanding of reversibility and continuity generalization. Two construct maps have been hypothesized for the first two dimensions whereas no construct map is available yet for the third dimension. In conducting a study for this article, we combined the two and focused therefore on investigating the progressions of students' understanding of "melting" (Mel) and "evaporation, boiling, and sublimation" (ESB). Table 1 presents a generic construct map for both Mel and ESB (and hence we use "MESB" as the label for the levels in this table).

As shown in Table 1, students in level 1 are assumed to be able to use the terminology (e.g., melting, evaporation, boiling, or sublimation) in an appropriate and relevant way, but do not show evidence of conservation. There are four subcategories used in coding level 1, depending on the nature of the evidence. In level 2, students do show evidence of conservation, but the evidence is incomplete and/or their understanding of conservation is incomplete. Again, four subcategories are used to give more detail. In level 3, students show evidence that they understand that both the identity and the weight of the substance are conserved in this type of phase change. In addition, level 0 is for responses that give no indication of understanding about the change, and level 0B indicates that there are missing data for this question.

TABLE 1
The Generic "Melting, Evaporation, Boiling and Sublimation" Construct Map

Level	*Description*	*Typical** *Scores*
MEBS3	Both Mass and Material are conserved.	3
MEBS2A	Mass conserved but Material is not conserved.	2
MEBS2B	Material conserved but Mass is not conserved.	2
MEBS2C	Mass conserved and Material is not addressed.	2
MEBS2D	Material is conserved and Mass is not addressed.	2
MEBS1A	Mass is not conserved and Material is not addressed.	1
MEBS1B	Material is not conserved and Mass is not addressed.	1
MEBS1C	Neither Material nor Mass is conserved.	1
MEBS1D	Something broadly relevant but that doesn't link to conservation of either.	1
MEBS0A	Neither Material nor Mass is addressed (and not 1D).	0
MEBS0B	No response, although they were asked.	0
MEBS0B	Not asked.	X

*Note that these scores may vary depending upon the circumstances, for example, if an item will allow no scores of 2, then the score of 3 would be reduced to 2.

Note that when the item does not prompt anything about either mass or material, then the response is automatically restricted to the "not explained" categories. Hence, if only Mass is addressed in the possible responses, then the codes can only come from {2C, 1B, 1D, 0}. Similarly if only Material is addressed, then the codes can only come from {2D, 1C, 1D, 0}. Note that if the item is open ended, it may only *ask* about one of these dimensions, but you might actually get responses that refer to the other (but they tend to be somewhat rare). In general, therefore, level 2 will be the upper limit for the responses we get to this type of item.

The concept of a construct map is an attempt to embody this developmental perspective on assessment of student achievement and growth (Masters, Adams, & Wilson 1990; Wilson 1990). A construct map is a well-thought-out and researched ordering of qualitatively different levels of performance (note that this does not imply that there is only one way for development to proceed, as there can be multiple categories in a level). Thus, a construct defines what is to be assessed in terms general enough to be interpretable across a curriculum but specific enough to guide the development of the other three building blocks. When the teaching objectives are linked to the construct, as will be outlined for Principle II, then it also defines what is to be taught (at a certain level of generality). Construct maps provide a way for large-scale assessments to be linked in a "principled" way to what students are learning in classrooms.

The idea of a construct map can be seen as a component of a learning progression, as described above. In this formulation, the learning progression is composed of a set of construct maps, each intertwined and dependent on the other. For example, one might see that, for the learning progression associated with teaching atomic-molecular theory in middle school science that we propose here, progress in this curriculum domain could be conceptualized as being built up from several strands, that is, construct maps in such areas as Properties of Objects, Density, Conservation and Change, Properties of Atoms and Molecules, and Molecular Theory of Macro Properties. This then corresponds to the conception of a *broad* learning progression, allowing individuals to follow different paths within it (across and within the construct maps), nevertheless, having discernible typical patterns (or "profiles" of progress) and allowing a broad characterization of general levels of sophistication (perhaps a profile rather than a single level) that can be useful to teachers and others involved in planning and reflection.

The approach assumes that, within a given curriculum, student response to the curriculum can be traced over a set of construct maps during 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 of the assessment system as well. Traditional testing practices—in standardized tests as well as in teacher-made tests—have long been criticized for over-sampling items that assess only basic levels of knowledge of content and ignoring more complex levels of understanding. Relying on constructs to describe what skills are to be assessed means that assessments focus on what is important, not only on what is easy to assess. Again, this reinforces the central instructional objectives of a course. In a large-scale assessment, the notion of a construct map will be more useful to the parties involved than simple number-correct scores or percentiles relative to some norming population, because it can help ensure that the results provide a certain level of diagnostic interpretation, as is so often requested.

The idea of using constructs also offers the possibility of gaining significant *efficiency* in assessment: Each new curriculum prides itself on bringing something new to the subject matter. In some cases, where quite radical innovations are made—for example, including a topic that has not hitherto been included in any school curriculum—new construct maps will have to be formulated. However, in the majority of cases innovations are interspersed into a common bedrock of curriculum. As the influence of national ("common") and state standards increases, this will become more true, and also easier to identify as such. Thus, we might expect innovative curricula to have one, or perhaps even two construct maps that do not overlap with typical curricula, but the remainder will form a fairly stable set of construct maps that will be common across many curricula. Given such flexibility, and the options that arise because the different strands can be intertwined in different ways, it can be seen that the approach is not prescriptive about the curriculum, it only requires synergy between the curriculum plan and a set of construct maps that reflect it.

Formative Assessments

As argued above, formative assessment requires a basis in a model of learning; the fact that *assessment for learning* is used widely as an alternative label to formative assessment indicates that a key principle here is that assessment is designed and used to directly further the learning of the student. Moreover, a strategy to "further" learning implies a need for a model of progress in learning that the teacher can use to compose and deploy assessment activities. Assessments constructed according to the BEAR approach can meet these requirements of formative assessment. In particular, the idea of a construct map can enhance the quality of formative work for it provides a tool to enable the teacher to move from the use (often only partly conscious or explicit) of inadequate models of progress to use of an explicit model that can be an object for reflection, exchange, and debate. These considerations can apply to many different teaching events, ranging from a brief informal discussion in the classroom to feedback from a formal terminal test.

Construct maps can provide an essential link between the teacher's understanding of the curriculum and how their teaching might work to achieve the goals of that curriculum. They can help make concrete the alignment of the two and, so, ensure that teaching that is consistent with the assessment framework is at the same time in line with the curriculum.

In the situation of a teacher designing learning work for a particular class on a particular day, a first decision is to select a task that will take the learning forward, that is, towards a fuller achievement of the curriculum's aims from the point already reached. A task will better serve this purpose if it is designed with the relevant construct map in mind. For example, if responses in a classroom dialogue indicate that the task is an easy one, then a task at a higher level of the construct is called for, whereas, if it is only producing evidence of confusion, then the teacher can select a task at a lower level. To choose a task that is both curriculum relevant and at the right level, a teacher has to judge the level that the class has reached; this estimate will be most accurate if the teacher has been interpreting work done so far on the curriculum topic in terms of the relevant construct map or maps.

Summative Assessments

The main difference between the situation for summative assessments and that for formative assessments is that the relevant grain-size of the construct maps will increase as the summative purpose becomes more removed from the classroom context. Hence it may be the case that several construct maps that were useful for formative assessments in the classroom will need to be combined in order to make their outcomes useful for summative purposes. For example, the several construct maps mentioned above as part of the topic of atomic-molecular theory (i.e., Properties of Objects, Density, Conservation and Change, Properties of Atoms and Molecules, and Molecular Theory of Macro Properties) may, for purposes of monitoring student growth over a part of the year, need to be combined into an overall Atomic-Molecular Theory construct map. This could involve the re-development of a new construct map at the aggregate level, or it might be carried out in a more top-down way, using less-demanding techniques to combine the construct maps (an example of such a technique would be the "described variable" approach used in the PISA tests [OECD, 2003]).

In addition, the scoring of the items may not need to be carried out in so detailed a fashion, and hence, the items themselves may involve fewer, or simpler, prompts for the students. In most other terms, summative assessment is similar to formative assessment. In particular the requirement for a relevant and validated model of learning is just as much needed here as it is in formative assessments, notwithstanding the coarseness of the interpretation that is to be made.

Principle II: Match Between Pedagogy and Assessment

The match between teaching and pedagogy in the BEAR Assessment System is established and maintained through two major parts of the system: construct maps (described above) and assessment tasks or activities (described in this section). The main motivation for the construct maps so far developed is that they serve as a framework for the assessments and as a method of making measurement possible. However, this second principle makes clear that the framework for the assessments and the framework for instruction must be one and the same, and, as shown in Figure 2, both must be strongly linked to the framework for the curriculum. This is not to imply that the needs of assessment must drive either the instruction or 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 which underlie the curriculum, whatever those aims are determined to be.

Using construct maps to structure both teaching and assessment is one way to make sure that the two are in alignment, at least at the planning level. One good way to achieve this is to develop both the teaching materials and the assessment tasks at the same time—adapting good teaching sequences to produce assessable responses and using evidence elicited by these responses to improve and enrich teaching activities. Doing so brings the richness and vibrancy of curriculum development into assessment and also brings the discipline and the explicit evidence of assessment data into the design of instruction.

Building Block 2: The Design of Tasks

The tasks 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 construct map and that the student responses can be coded into categories that correspond to levels of the construct map.

As was emphasized above, a task as such is neither formative nor summative: it is the way that it is used—in particular, the way in which learners' responses are interpreted and used—that constitutes the distinction. Insofar as tasks for assessment are related to teaching usages, a variety of different task types may be used, so providing for the requirements of a variety of teaching situations but also reflecting the different learning contexts of those situations. There will therefore be a variety of different types of assessment tasks designed primarily for formative use, just as there will be a variety in those designed primarily for summative use. Insofar as such tasks are used for the latter purpose, they may be open ended, requiring students to fully explain their responses in order to achieve a high score, or they may be multiple-choice (e.g., Briggs, Alonzo, Schwab & Wilson, 2006), or they may be computer delivered and scored (e.g., Scalise & Wilson, 2011), freeing teachers from having to laboriously hand-score all of the student work. 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. In the BEAR Assessment System, this tension becomes something to design for and control—both types of item may be included and the design task is to judge the optimum mix in the light of the given context and constraints. For any summative test, feedback on the results can involve pupils and so serve a formative purpose as well. It is also possible to use both open-ended and multiple-choice items in on-going teaching situations: for example, a class may be presented with a multiple-choice question and asked to discuss it briefly and then to vote on the options. These votes will help to indicate to the teachers which issues, if any, need more discussion time.

Based on the work of Smith et al. (2007), a group of teachers from San Francisco Unified School District developed a set of items to investigate, in summative mode, the performance of

ITEM 219

Carmen wants to know what happens to the mass of something when matter changes from one form to another. She puts three ice cubes in a sealed bag and records the mass of ice in the bag to investigate this. Once the ice cubes have melted, she records the mass of the water in the bag. Which of the following best describes the result?

a) The mass of the water in the bag will be less than the mass of the ice in the bag.
b) The mass of the water in the bag will be more than the mass of the ice cubes in the bag.
c) The mass of the water in the bag will be the same as the mass of the ice cubes in the bag.

Describe your thinking. Provide an explanation for your answer.

FIGURE 7 A dual multiple-choice and open-ended item, with the open-ended item hypothesized as an indicator for the full range of levels X to 3 of Table 1.

their students on the Melting and ESB constructs. Given that our work aimed only to illustrate our "road map" approach, the items used here were selected from those used by the teachers and from other sources in the literature, and were not generated for the particular purpose of our work. An example of an open-ended item they hypothesized to be likely to generate responses at levels 0 to 3 (Table 1) of the MESB construct map is shown in Figure 7 (see Appendix B for the text of all of the items). They generated quite a few items initially. Eventually, after considerable qualitative and quantitative analyses, a smaller set remained: 1 multiple-choice and 4 open ended for Melting and 2 multiple-choice and 4 open ended for ESB. (Note, for ESB only evaporation items remained, hence, we should consider this realization to be a measure of Evaporation only.) The example in Figure 7 was designed to elicit first a multiple choice response, which was scored according to the choice made, and then a constructed response, scored according to the description and explanation offered.

Creating the developmental construct maps is not a trivial task. The Smith et al. (2007) paper was based on a review of many years of research by many researchers, and itself took almost a year to write—our subsequent work to get to this point at which we have hypothesized construct maps has spanned approximately 12 months. But having succeeded in adapting this approach to a given curriculum, educators will be well situated to address many of the issues raised in the introduction to this article.

Formative Assessments

For formative purposes, the alignment between pedagogy and assessment has to be effective at the level of classroom interaction, and that is the point at which the nature of the assessment tasks becomes crucial. Assessment tasks need to reflect the range and styles of the instructional practices implied by the curriculum; they must have a place in the "rhythm" of the instruction, built in as part of the constant interaction that is essential to ensure that the teacher and the learner are mutually and closely involved in a common purpose. They work together to build a shared understanding of each of the steps along the road to achieving that purpose.

It should be clear that from this perspective, formative assessment is a more complete development of the idea of *embedded assessment*; the latter only requires that the assessments be a part of the regular classroom activities. But for assessment to become fully and meaningfully embedded in the teaching and learning process, the assessment must be linked to the specific curriculum in use, that is, it must be curriculum dependent, not curriculum independent—in contrast to the situation that is the premise in many high stakes testing situations (Wolf & Reardon, 1996). 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 & Baker, 1996).

Summative Assessments

In large-scale testing situations, the mix of task types will often be 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 some assessment tasks might be valued because of their perceived high validity, they may not yield enough information to accurately estimate each examinee's proficiency level given limited testing and scoring time. In particular, multiple-choice items, which require less time to answer and which can be scored by machine rather than by human raters, may be more heavily used, to increase the reliability of a summative test. Special care will be needed to ensure that this does not damage the validity of the interpretations to be made based on the outcomes. Another way in which the items' design may differ is that summative assessment items may emphasize more global questions; an example of such a question could be "Explain what happens when a substance changes state?" Such a question covers several different construct maps in the framework that would comprise component 4 of Figure 3 and, hence, would not necessarily be included when the perspective was at the grain-size of the individual construct map. That said, this might still be an interesting question to ask in a formative setting, as its generality and vagueness provide the opportunity for wide-ranging discussions about student responses: it could be used at the start of work on a topic to give the teacher a general overview of the starting levels of a class's understanding.

Principle III: Management by Teachers

For information from the assessment tasks and the BEAR analysis to be useful both for teachers and students, it must be couched in terms that are directly related to the instructional goals behind the construct maps. This is clearly essential if the summative result is to be used to guide a review of a learner's progress, and/or to serve as a guide to decisions about the next steps in a learner's work, and/or to indicate changes in the practices and management of the teaching that may be required. These purposes require the use of classroom evidence even if, as purposes specific to large-scale summative assessment, they are given priority. In such large-scale assessments, open-ended tasks, if used, must be able to be quickly, readily, and reliably coded and scored. The categories into which they are coded 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 how item outcomes relate to how students actually respond to an item leads to the third building block.

Building Block 3: The Outcome Space

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

The information from assessment opportunities must be couched in terms that are directly interpretable with respect to the instructional goals of the constructs. Moreover, this must be done in a way that is intellectually and practically efficient. Scoring guides designed to meet these two criteria can serve as a practical definition for a construct by describing the performance criteria necessary to achieve each score level of the construct. 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) to all users.

Traditional multiple choice items are, of course, based on an implicit scoring guide—one option is correct, all the others are 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. The outcome space for the multiple-choice part of the Melting item is shown in Figure 8, where the levels for each option are shown.

The open-ended part of the Melting item needs a more detailed outcome space.[9] It is shown in Figure 9. Note that some of the levels have more than one code within them. Often it will be the case that an item generates responses that have educationally useful subcategories within particular levels. The codes 2A, 2B, and 2C are examples of these. Generally, these subcategories are used in the qualitative analysis of the data but are often ignored in the quantitative analysis (where all of these three would be collapsed to "2").

A. Multiple-Choice

Response	New Code
Option C	2C
Option A/B	1A
No Response	0

FIGURE 8 The outcome space for the multiple-choice part of the item shown in Figure 7.

[9] The full set of scoring guides for the 11 items is shown in Appendix B.

B. Open-ended

Description	*Typical Student Responses*	*Code*
Both same mass (or weight) and same material explicitly	*"(Students chose c) Mass will be the same because they are the same substance but in a different form."* *"(students chose c) The mass of the matter will be the same as the ice cubes' mass because only its state of matter changed and not the molecules themselves."*	3
Both different (larger or smaller) mass (or weight) and same material explicitly	*"(Students chose b) The water will have more mass because the water will have more mass if turned into liquid."* *"(Students chose a) I think the mass of the water in the bag will be less because the pressure changes. It's still the same amount of water, but the molecules move."*	2B
Same mass (or weight) but not explicit about same material	*"(Students chose c) I think that the mass will stay the same because matter cannot be destroyed or created and if nothing can escape the bag it will stay the same in mass at least."* *"(Students chose c) Even if matter changes to another, the mass will still be the same because there are the same number of molecules in there."*	2C
Same mass but not same material	*"(Students chose c) It stays the same because you put an amount in there & just because the molecules change doesn't mean the amount has."*	2A
Different (larger or smaller) mass (or weight) but not explicit about same material	*"(Students chose b) Mass is space taken up. So I think it's b, since water fills up the bag but the ice cubes take less mass since it's bundled up into."* *"(Students chose a) Ice is solid and weighs more than liquid."*	1A
Any response that indicates the term *melting* incorrectly used/I don't know/Off-topic response	*"I don't know why."* *"I guessed."*	0
Blank		0

FIGURE 9 The outcome space for the open-ended part of the item shown in Figure 7.

This item, and others we have used from the Smith et al. paper, satisfies the design principles set out above. The items seek to generate student responses in the construct maps to which they were attached; this is their most important design feature (Wilson, 2005). However, as will be explained below, they do not conform to a comprehensive scheme of design principles, as might a set of items designed according to the BAS principles. This should not be surprising, as they were generated to illustrate the results from a wide-ranging base in the research literature, and were not intended to exhibit such consistency.

Formative Assessments

When considering formative uses of the outcome space, it is useful to think of the metaphor of a stream of activity in which teacher-student and student-student interactions engage the students in actively developing their own learning. The teacher sets the stream flowing, by the choice and explanation of the task, and then dips frequently into this stream of learning to evaluate student progress, interacting closely and frequently in order to steer, challenge, and move on the dialogue so that the students' learning makes authentic progress. In this metaphor, the assessment interaction becomes an intrinsic and essential part of the teaching and learning process. This is the core meaning of formative assessment.

In order to conduct a learning dialogue in this way, a teacher has to be able to locate responses in relation to the construct maps, so that the point reached can be evaluated, the next possible step foreseen, and the challenges and/or steering interventions optimally framed. In this view of formative learning, traditional "scoring," in the sense of merely assigning numbers to students' work as a reflection of a judgment of quality, is irrelevant; indeed it can do harm in moving students into an ego-involved attitude rather than a task-involved approach (Butler, 1987; Dweck, 2000). The existence of the construct map and outcome space means that the estimates of student location on the Wright map (see next building block) can be embedded in a web of interpretation that transcends the traditional "score" on a test. (Note that it is not the mere application of item response theory or Rasch modeling that give this advantage, it is the combination of those measurement models with the rest of the BAS approach that makes this possible and practicable.)

The difficulty that teachers face in all of these activities is the difficulty of interpreting the students' contributions, particularly where these are unexpected or partially correct, in order to locate them in a progress scheme so that they can respond, sometimes at very short notice, to adapt or modify the next teaching action to meet the needs thus exposed. What the teacher requires are ways to interpret student responses that can give direct help in framing feedback comments appropriate to what the students' have said or written and that give appropriate guidance to help them progress. Thus, scoring guides that set out qualitative indices to identify the location of any particular response along a construct map can be an essential part of making construct maps helpfully applicable.

Summative Assessments

Teachers need summative results about their students, at a variety of grain sizes, to make both group-level decisions (about such questions as "Have enough of my students improved to a level such that the next topic or module can be readily learned by them."), and individual-level decisions (about such questions as "Has Tommy done well in Science this year?"). In addition, educational administrators at a variety of administrative levels, also need summative results (regarding questions ranging from "Has Miss White's class learned to the same level as Mr. Brown's class?" to "Have the students in grade 4 in my school district learned as much about science as the students who were in grade 4 last year?"). Hence, the need of the users of summative assessments with these different purposes is, primarily, that the results be communicated in a way that is both meaningful and reasonably efficient. The graphical basis for the Wright maps allows for a variety of output formats that can convey large amounts of information and link them visually to interpretative categories aligned with the levels of the construct map, thus making interpretation more accessible to all.

Principle IV: High Quality Evidence

Technical issues of reliability, validity, fairness, consistency, and bias can quickly sink any attempt to measure along a construct map 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) map student performances onto the construct maps, (b) map items and raters onto the construct map, (c) establish uniform levels of system functioning in terms of quality control using reliability indices, and (d) establish validity evidence using a variety of evidentiary criteria (e.g., Wilson, 2005, Chapter 8). While such considerations can become very technical, it is sufficient to keep in mind that the traditional elements of assessment standardization, such as validity and reliability studies and bias and equity studies, must be carried out to satisfy quality control and ensure that evidence can be relied upon (e.g., as in AERA/APA/NCME, 1999).

Building Block 4: Wright Maps

A Wright map is an empirically-based version of a construct map: It represents the principle of high quality evidence. Wright maps are graphic and empirical representations of a construct, showing how it unfolds or evolves in terms of increasingly sophisticated student performances. It is based on the analysis of student responses to the tasks, moderated through a statistical model that expresses one's expectations about the underlying construct map (see Wilson, 2005, Chapter 6, for a description of the technical basis for the Wright map).

An illustrative Wright map is shown in Figure 10. It is based on students' responses to the 11 items developed by the SFUSD teachers mentioned above. The items were piloted in a paper-and-pencil format in June 2008. Six hundred eighty-six grade 8 students from eight schools in the San Francisco Unified school district took the test at the end of the school year. The sample consists of 344 females and 342 males. Students responding in Spanish were dropped from the sample and thus the final sample size was reduced to 665.

In Figure 10, the columns labeled "Student" shows the distribution of the locations, in terms of their overall scores, of the students for each of the Melting and Evaporation constructs. Each "X" represents the location of a small number of student's along the construct map (in this case, 5.5 students). Locations of the item score levels (i.e., for the individual items) are shown in the columns to the right of each "Student" column—these are the points where half of the students below this location get to at least that score. Thus, for item 219OE, "1" is the overall score location for which half of the students scored below 1 and half scored 1 or above. In subsequent use, we would predict that students at this location would get at least this score 50% of the time. Likewise, "2" is the location for which half of the students scored below 2 and half scored 2 or above. A similar interpretation holds for the multiple-choice responses for item 219 ("219MC"). As one might expect, the attainment of a level 1 response for the multiple-choice part of item

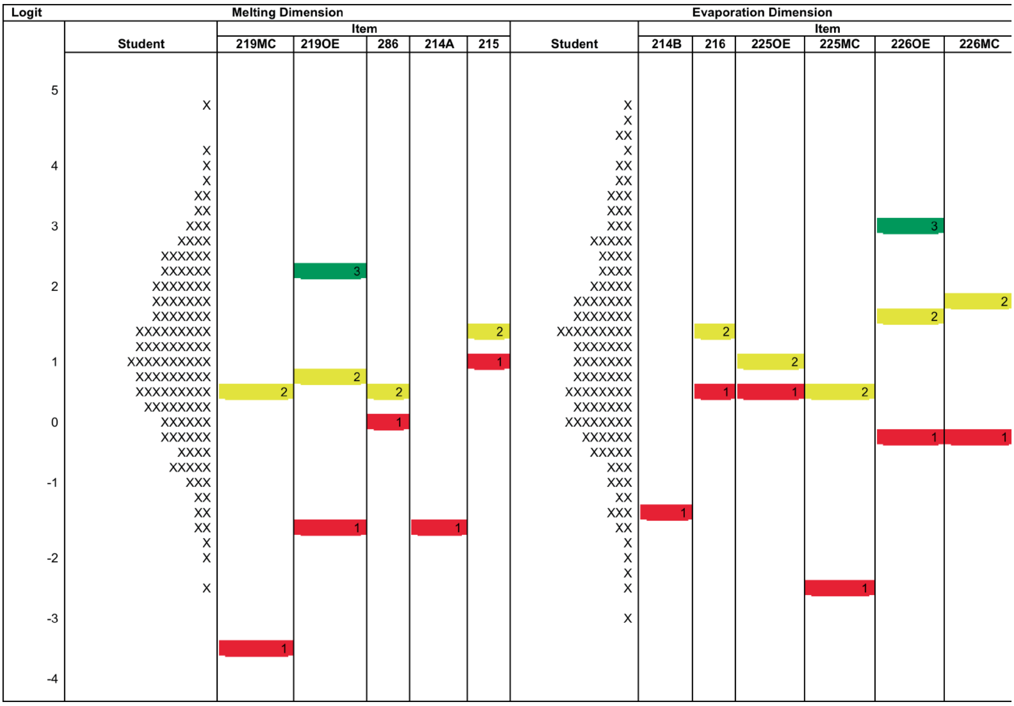


FIGURE 10 A Wright map of a construct map related to atomic molecular theory.

219 is easier than for the open-ended part of the item. But, the open-ended portion of the item relates to a higher portion of the scale—that is, the higher scores on the open-ended portion are higher than the highest score possible on the multiple-choice portion.[10] The scale used for the numbers on the left are not raw scores, but are in units called "logits," or the log of the odds. A multi-dimensional Rasch modeling approach is used to calibrate the maps for use in the BEAR Assessment System (see Adams, Wilson, & Wang, 1997, for the specifics of this model).[11] The two construct maps, Melting and Evaporation, have been calibrated onto two dimensions[12] to create this map, and the correlation between those dimensions was found to be 0.68.

A key feature of these maps is that both students and tasks can be located on the same scale, giving student performances 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.

[10] Note that this relationship between the multiple-choice part of item 226 (226MC) and the open-ended part of item 226 (226OE) is not analogous to the multiple-choice and open-ended parts of items 219 and 225 (see Appendix B).

[11] The reliabilities of the Melting and ESB item sets were found to be 0.73 and 0.75, respectively.

[12] The metrics of these two dimensions have been aligned using a dimension-aligning technique (Yao, 2010).

Formative Assessments

Wright maps can be powerful tools to guide the arc of learning through formative interactions. They can be used to identify, record, and track student progress and to illustrate the skills that students have mastered and those that the students are working on. By placing students' performance on the continuum defined by the map, teachers and others can interpret student progress with respect to the standards that are inherent in the construct maps.

All of the above applies to both written and oral forms of learning dialogue, but it assumes that the teacher is in control of the learning progress. However, productive formative assessment should also incorporate self- and peer-assessment by students as a part of classroom practice, for it is in such activities that students take more responsibility for their own learning and so are helped to become independent learners in the future, as well as improving their attainments in the short term. Here again interactive feedback is essential, whether through the feedback of self-reflection or through feedback from peers. Students can only develop the skills of self- and peer-assessment if they can generate such feedback, and for this they must have a clear understanding both of the targets towards which their activity should be aimed and of the criteria of quality against which to make their assessment judgments. Scoring guides can meet this need, but only if they are shared with, and understood by, the students. The value of such sharing has been attested both in the United Kingdom (Black et al., 2003; Wiliam et al., 2004), and in the experience of the use of the BEAR Assessment System (Roberts & Sipusic, 1999). However, the ways in which this might be done will differ according to the ages of the pupils: for younger pupils, the language will have to be more simple, and visual devices will be more valuable for younger pupils than for older ones (see, e.g., Harrison & Howard, 2009).

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, the BEAR Center has also developed software for teachers to use to generate the maps. This software, which we call *ConstructMap* (Kennedy, Wilson, Draney, & Tutunciyan, 2007), allows users to enter the responses given by students to assessments, and then to map the performance of groups of students, either at a particular time or over a period of time.

Summative Assessments

Wright maps can be very useful in large-scale assessments, providing information that is not readily available through numerical score averages and other traditional summary information—they are used extensively, for example, in reporting on the PISA assessments (OECD, 2005). These maps have at least two advantages over the traditional method of reporting student performance as total scores or percentages. First, it makes possible the interpretation, with respect to the standards that are inherent in the construct maps, of 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. It should be borne in mind however that a Wright map derived from (say) a particular school grade group would not necessarily give information consistent with that from a different sample of students. For example, the relative difficulties of any two items might differ because the learning experiences of the two samples have been different. At the same time, however, evidence of such differences would be useful indicators for the teachers involved.

Linking Formative and Summative Assessments in the BAS

The formative and the summative are two different purposes assessments can be designed to serve. Some take the view that these purposes are so different that they cannot use the same instruments and cannot be interrelated in a common system of assessment. An alternative view is that, whilst some tasks may be useful for one purpose only, many can serve both purposes, the differentiation being made in the interpretation of the responses and the use of those interpretations.

Assessments generated and used within a school are important because they can serve the purposes of advice and guidance to students for formative purposes such as improving present achievements, and/or for summative purposes such as making choices about further stages of their study. Achieving a high standard of validity and reliability is important for these classroom purposes, as it is for large-scale assessments, although reliability is less crucial for formative work as weakness in this respect will usually come to light quickly and can be corrected; the nature of the evidence will differ, given the different settings, it may be interpreted differently according to whether the purpose is formative or summative (Wiliam & Black, 1996). In fact, teachers have opportunities to achieve far better standards of validity and reliability for their in-classroom assessments than can be achieved for national or state tests because they can generate richer and more meaningful pieces of evidence compared to what an externally imposed system can reveal. However, whilst this is possible in principle, it might require a program of professional development to help many teachers develop the insight and the skills required. Examples of such work are described in Black et al. (2010, 2011).

Such alignment of formative and summative uses is clearly a desirable outcome, one that the BEAR system supports, in part because it serves as a framework for both formative and summative and, more directly, because it allows for information from one to be used as part of the other. Thus, for example, the difference indicated in Figure 10 between items 225MC and 226MC, and the fact that the difference between the corresponding open-ended items is far smaller, would call for further consideration by any teacher. The outcomes might either indicate significant problems in the teaching work or deficiencies in one or more of the items involved. Alignment is also desirable because it makes efficient use of the time devoted to assessment, and because both formative and summative work ought to be closely linked to the aims of the curriculum and be grounded in a common view of learning. There ought to be synergy and support between the two rather than the tension teachers experience when external pressures seem to be in conflict with their own beliefs about the teaching that best serves the learning of their students.

The BEAR system is not merely a multipurpose tool that can be adapted for one assessment purpose or the other. Its emphasis on alignment of curriculum and assessment is consistent with the need for such alignment in both formative and summative assessment. The argument of Resnick and Resnick (1992) that “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” (p. 59) makes clear that construct maps that embody the aims of teaching (e.g. “standards”) can guide formative assessment to meet those “real goals of educational reform.”

For formative purposes, such alignment must be at a level of detail that can guide day-to-day activities, and must address the complex and varied routes that students follow in their progress in learning. Here the BEAR system can be used as a basis for organizing the thinking and the tools needed by teachers to make this both possible and practicable. For summative purposes, the alignment will be at a higher level of grain size, the coarseness of the grain size depending on how far the planned usage of the outcomes is from day-to-day classroom usage. For example, a test used for a formative purpose can provide a variety of pieces of evidence for the purpose of checking progress at an interim stage in the teaching of a topic, evidence that could be used to identify both strengths and shortcomings at the classroom level. The same evidence might also be aggregated to provide a score, or a profile, which can be recorded alongside other evidence obtained on other occasions to guide decisions about future choices or assignments. Finally, such shared-purpose tasks can be used in two ways in externally generated high-stakes tests. They can be used (a) to generate school-based outcomes that can be a part of an external assessment system and (b) as models for tasks to be used in external tests. Use of either or both of these strategies would go a long way toward making an accountability system work more coherently and more successfully (NRC, 2005).

The common basis for both purposes would be the construct map, particularly where this is expressed in the form of a Wright map. It is also worth noting that the generation of a construct map may be served by the collection and interpretation of data from tests, whether designed and used for formative or for summative purposes, and that after these data have been combined, through their different contributions, in the shaping of the construct map and its development into a Wright map, the outcome can be a common guide to the shaping of both formative and summative assessment activities.

5. DISCUSSION AND CONCLUSION

We have set out in Section 1 a model that represents the key features of classroom instruction that, through its exploration of the concepts of formative and summative assessment, helps to clarify the nature of the ways in which assessments should serve these two purposes. This then shows that a coherent and supportive relationship between the two uses of assessment, that is, to serve both formative and summative purposes, can only be achieved if the same underlying map of progression in learning is adopted for both assessment *for* learning and assessment *of* learning. It is worth noting here that, in their exploration of the application of activity theory to formative work, Black and Wiliam (2006) pointed out that the formative "tools" they have developed with schools would lack substance unless they could be backed by a scheme of progression grounded in evaluation of test items, for which the BEAR research provides a model (Wilson & Sloane, 2000).

Section 2 has taken the argument further by pointing out that a relationship this coherent and supportive can only be secured if the underlying map is in harmony with the curriculum requirements teachers choose or are required by their communities and states to adopt as the overall framework of aims within which their teaching plans are drawn up. It is arguable that any externally imposed curriculum should be less detailed than this approach may require, but it is hard to see how a summative testing strategy can achieve validity unless it is based on a detailed scheme and reflects one way, or several possible ways, of interpreting the curriculum in terms of progression in learning.

The challenge then is to formulate ways to compose such a map and to evaluate and refine it in practice. The discussion in Section 3 explains the first phase, that is, the formulation of a map, by a detailed discussion of one particular example. This discussion fulfilled its first aim of setting up a specific example to be used to exemplify the argument of Section 4, but it also showed that research studies can only guide, rather than specify in detail, this formulation. It follows that by drawing on different sources, or by tuning such a map to the different contexts, notably of language and culture, within which any scheme of instruction has to be implemented, different maps might emerge; these differences might be legitimate if each map were the optimum for the context for which it had been framed.

Section 4 has then described the BEAR method and shows the results that were established from its application to the analysis of a particular set of questions exploring pupils' understanding of Changes of State. The results illustrated in a Wright map that this produces derive from the questions and the level assignments used to analyze the responses. In general, the assumptions about the approach to a learning progression adopted by a teacher, in the light of any chosen or prescribed curriculum, would determine or at least heavily bias the teaching approach and any assessment instruments used, and the response analysis schemes would all reflect these choices. Thus a road map will provide results about the outcomes of what the teacher has decided to do, and it would reveal problems and inconsistencies in the assumptions about optimum progression entailed by those assumptions.

In systems in which teachers might be permitted to be flexible in their interpretation of a curriculum specified only in broad terms, they would need means to both develop and evaluate the detailed decisions about learning progressions that followed from these interpretations, and would need ways to check these decisions. The approach developed in this article supported by the software tools used here, would serve this need. This resource would be important in any system that took to heart the argument (e.g. by the UK Assessment Reform Group, 2006) that teachers should play a significant role in the summative assessments of their pupils for all types of purpose, whether within schools or across schools.

A different road map, emerging from a different complex of curriculum, teaching, questions, and response analyses would reveal different information. The two different maps may have a complex relationship (they may or may not be commensurate, for example). The two approaches would be alternatives within the topic specified in one box of a road map of the type shown in Figure 3. The results for two (say) groups of students taught according to two different schemes that focused on the same box might then be analyzed separately and compared. Any differences might then point to amendments that might need to be made in relation to the box topic, and/or might expose adaptations that might have to be made in other, dependent boxes and, so, make the work coherent overall.

Whilst the example discussed here is about a science topic, it would be wrong to infer that the system can only be implemented in subjects for which assessment scoring is mainly analytic in nature. For example, for a test, or portfolio of classroom tasks, in (say) the writing of essays in English, a holistic approach to each component may be the only appropriate one. What would normally be required, in any system, would be a level or ranking assignment to any individual component and some rule for aggregation to arrive at an overall level or grade for the whole of the test or portfolio. These data could then be used within the BEAR system to construct a Wright map.

ACKNOWLEDGMENTS

This article is jointly authored; authors' names are listed in alphabetical order. We would like to note that we are indebted to the authors of Smith, Wiser, Anderson & Krajcik (2006) for laying out the initial version of the learning progression for our own work in this article.

REFERENCES

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

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Alexander, R. (2006). *Towards dialogic teaching: Rethinking classroom talk* (3rd ed.). Cambridge, United Kingdom: Dialogos.

Andersson, B. (1990). Pupils' conceptions of matter and its transformations (ages 12–16), *Studies in Science Education, 18*, 53–85

Applebee, A. N., Langer, J. A., Nystrand, M. & Gamoran, A. (2003). Discussion based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal, 40*(3), 685–730.

Assessment Reform Group. (2006). *The role of teachers in the assessment of learning*. London, UK: Institute of Education.

Bar, V., & Galili, I. (1994). Stages of children's views about evaporation. *International Journal of Science Education, 16*(2), 157–174.

Bar, V., & Travis, A. S. (1991). Children's views concerning phase changes. *Journal of Research in Science Teaching, 23*(4), 363–382.

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

Black, P., Harrison, C., Hodgen, J., Marshall, M., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education, 17*(2), 215–232.

Black, P., Harrison, C., Hodgen, J., Marshall, M. & Serret, N. (in press). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education*.

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–71.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In John Gardner (Ed.), *Assessment and Learning* (pp. 143–181). London, United Kingdom: Sage.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31.

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

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

Bruner, J. S. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology, 79*(4), 474–482.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance, *British Journal of Educational Psychology, 58*, 1–14.

Butler, R., & Neuman, O. (1995). Effects of task and ego-achievement goals on help-seeking behaviours and attitudes. *Journal of Educational Psychology, 87*(2), 261–271.

Chang, J.-Y. (1999). Teachers college students' conceptions about evaporation, condensation, and boiling. *Science Education, 83*(5), 511–526.

Claesgens, J., Scalise, K., Draney, K., Wilson, M., & Stacy, A. (2002, April). *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.

Costu, B., & Ayas, A. (2005). Evaporation in different liquids: Secondary students' conceptions. *Research in Science and Technological Education, 23*(1), 75–97.

Dweck, C. S. (2000). *Self-theories: their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Glaser, Robert. (1990). The reemergence of learning theory within instructional research. *American Psychologist, 45*, 29–39.

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.

Hallam, S., & Ireson, J. (1999). Pedagogy in the secondary school. In P. Mortimore (Ed.), *Understanding pedagogy and its impact on learning*. London, United Kingdom: Paul Chapman.

Harrison, C., & Howard, S. (2009). *Inside the primary black box*. London, United Kingdom: GL Assessment.

Johnson, P. (1998a). Children's understanding of changes of state involving the gas state, Part 1: Boiling water and the particle theory. *International Journal of Science Education, 20*(5), 567–583.

Johnson, P. (1998b). Children's understanding of changes of state involving the gas state, Part 2: Evaporation and condensation below boiling point. *International Journal of Science Education, 20*(6), 695–709.

Johnson, D. W., Johnson, R. T. & Stanne, M. B. (2000). *Co-operative learning methods: A meta-analysis*. Retrieved from: http://www.co-operation.org/pages/cl-methods.htm

Johnson, P., & Papageorgiou, G. (2010). Rethinking the introduction of particle theory: A substance-based framework. *Journal of Research in Science Teaching, 47*(2), 130–150

Kennedy, C. A., Wilson, M., Draney, K., & Tutunciyan, S. (2007). *GradeMap* (Version 4.2) [Computer software]. BEAR Center. Berkeley: University of California.

Klenowski, V. (2009). Assessment for learning re-visited: An Asia-Pacific perspective. *Assessment in Education, 16*(3), 263–268.

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, IL: University of Chicago Press.

Masters, G., & Forster, M. (1996). *Progress maps. Assessment resource kit*. Victoria, Australia: Commonwealth of Australia.

Masters, G. N., Adams, R. A., & Wilson, M. (1990). Charting student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary volume 2* (pp. 628–634). Oxford, United Kingdom: Pergamon Press.

Mercer, N., Dawes, L., Wegerif, R. & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal, 30*(3), 359–377.

Meyer, J., & Land, R. (2003). *Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines*. Edinburgh, UK: ETL Occasional Report 4. School of Education, University of Edinburgh.

Nakhleh, M. B., Samarapungavan, A., & Sablam, Y. (2005). Middle school students' beliefs about matter. *Journal of Research in Science Teaching, 42*(5), 581–612.

National Research Council (NRC). (1999). *How people learn: brain, mind, experience, and school*. In J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). Washington, DC: National Academy Press.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. In J. Pellegrino, N. Chudowsky, & R. Glaser, (Eds.), Washington, DC: National Academy Press.

National Research Council (NRC). (2005). *Systems for state science assessment*. In M. Wilson & M. Bertenthal, (Eds.), Washington, DC: National Academy Press.

OECD. (2003). *The PISA 2003 assessment framework*. Paris, France: OECD.

OECD. (2005). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: OECD.

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

Paik, S-I., Kim, H-N., Cho, B-K., & Park, J-W. (2004). K–8th grade Korean students' conceptions of "changes of state" and "conditions for changes of state." *International Journal of Science Education, 26*(2), 207–224.

Papageorgiou, G., & Johnson, P. (2005). Do particle ideas help or hinder pupils' understanding of particle phenomena. *International Journal of Science Education, 27*(11), 1299–1317.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field. *Assessment in Education: Principles, Policy and Practice, 5*(1), 85–102.

Prain, V., Tytler, R., & Peterson, S. (2009). Multiple representations in learning about evaporation. *International Journal of Science Education, 31*(6), 787–808.

Ross, K., & Law, E. (2003). Children's naive ideas about melting and freezing. *School Science Review, 85*(311), 99–102.

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*. Boston, MA: Kluwer.

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

Scalise, K., & Wilson, M. (2011). The nature of assessment systems to support effective use of evidence through technology. *E-Learning and Digital Media, 8*(2), 121–132.

Shawn, Y. S., Delgado, C., & Krajcik, J. S. (2010). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching, 47*, 687–715.

Shayer, M. (2003). Not just Piaget; not just Vygotsky, and certainly not Vygotsky as *alternative* to Piaget. *Learning and Instruction*, 13(5), 465–485.

Shayer, M., & Adey, P. (1981). *Towards a science of science teaching*. London: Heinemann.

Smith, F., Hardman, F., Wall, K., & Mroz, M. (2004). Interactive whole class teaching in the national literacy and numeracy strategies. *British Educational Research Journal, 30*(3), 395–412.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*,(1 & 2), (Entire issue).

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: What works best and issues for development*. Oxford University Centre for Educational Development, Report commissioned by the QCA. Retrieved from http://www.education.ox.ac.uk/assessment/uploaded/2009_03-Review_of_teacher_assessment-QCA.pdf

Stavy, R. (1990). Children's conceptions of the changes in the state of matter: From liquid (or solid) to gas. *Journal of Research in Science Teaching, 27*(3), 247–286.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Tytler, R. (2000). A comparison of year 1 and year 6 students' conceptions of evaporation and condensation: Dimensions of conceptual progression. *International Journal of Science Education, 22*(5), 447–467.

Tytler, R., & Peterson, S. (2005). A longitudinal study of children's developing knowledge and reasoning in science. *Research in Science Education, 35*, 63–98.

Varelas, M., Papas, C. C., & Rife, A. (2006). Exploring the role of intertextuality in concept construction: Urban second graders make sense of evaporation, boiling and condensation. *Journal of Research in Science Teaching, 42*(7), 637–666.

Webb, N. M., Nemer, K. M., & Ing, M. (2006). Small-group reflections: Parallels between teacher discourse and student behavior in peer-directed groups. *The Journal of the Learning Sciences, 15*(1), 63–119.

Wiliam, D., & Black, P. (1996). Meanings and consequences : A basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal, 22*(5) 537–548.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49–65.

Wilson, M. (1990). Measurement of developmental levels. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary volume 2* (pp. 628–634). Oxford, UK: Pergamon Press.

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, IL: University of Chicago Press.

Wilson, M. (2004b). A perspective on current trends in assessment and accountability: Degrees of coherence. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education, Part II*. Chicago, IL: University of Chicago Press.

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

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education, Part II*. Chicago, IL: University of Chicago Press.

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

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

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

Wiske, M. S. (1999). What is teaching for understanding? In J. Leach & B. Moon (Eds.), *Learners and Pedagogy*. London, United Kingdom: Chapman.

Wolf, D., & 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 I*. Chicago, IL: University of Chicago Press.

Yao, S-Y. (2010). *Aligning the melting and evaporation dimensions: Results of the theta and delta methods*. BEAR Research Report. Berkeley: University of California, Berkeley.

APPENDIX A

TABLE A1
Sample, Data and Analysis Details for the Sixteen Studies Quoted Data Described or Quoted in Andersson (1990) and Shayer & Adey (1981) are not Included in this Table

	Sample			*Intervention or*		
Author(s)	*Ages*	*Numbers*	*Type*	*curriculum linked*	*Data*	*Data analysis*
Bar & Galili (1994)	7 5–11 10–15	293 105 152	CS	CL	Interviews OE questions Interviews	Categories % by age
Bar & Travis (1991)	6–12	80	CS	Int	Interviews MC test OE questions	Categories % by age Factor analysis
Chang (1999)	MathSci No-Sci Litle-Sci	73 102 69	CS by college course	Int Demonstrations with 6 tasks	Written accounts	Categories % by age
Costu & Ayas (2005)	15 16 17	107 116 90	CS	No	OE questions 12 pupil interviews	Categories % by age
Johnson (1998a) Johnson (1998b)	11–14	33	L	Int: 4 teaching units	Questionnaire 1st unit: interviews 5 between & after units	Categories. Tracks changes over the years. Cross links findings & particle model beliefs.
Johnson & Papagiorgiou (2010)	9–10	30 in school (a) 15 school (b)	Review and new data	CL (a) did classes for 4 weeks; (b) for 12	(a) Open interviews (b) interviews about particles	Categories of topics: results of two groups compared
Papagiorgiou & Johnson (2005)	10–11	(a) 20 (b) 20	Experiment/control	Int (a) 12 week lessons; (b) 10 week lessons; only (a) with particles	Interviews before and after: 12 in each of the classes	Categories - compare numbers experiment with control
Nakhleh et al. (2005)	13	9	One age	No	Interviews OE questions	Categories % by age

(*Continued*)

TABLE A1
(Continued)

	Sample					
Author(s)	*Ages*	*Numbers*	*Type*	*Intervention or curriculum linked*	*Data*	*Data analysis*
Paik et al. (2004)	6–14	5 at each of 5 ages	CS	CL. Activities from text-books	Interviews before and after activities	Transcripts & field notes Categories and % by age
Prain et al. (2009)	11	9 from 3 classes	One age	Int: Class activities designed	Interviews and class recordings and later follow-up interviews	General account & case studies of three children
Tytler (2000)	6/7 11/12	28 28	CS	Int: Group tasks & inter-group reports	Record talk re tasks & pupils' drawings & questions about tasks.	Categories in talk & drawings and written work; %'s in these by age group
Tytler & Peterson (2005)	5–8	12	L	CL. Lessons designed with teachers'	Interview all twice per year over 4 years plus class recordings	Qualitative accounts
Ross & Law (2003)	12, 14,16,17	4,9,4,4	CS	No	Group interviews OE questionnaire	Categories
Stavy (1990)	9–15	20 at each of 6 ages	CS	Int: Demonstrations of real phenomena	Interviews about the demos.	Categories % by age
Varelas et al. (2006)	6	26 in one class	One age	CL. Linked to class teaching	Videos of lessons	Textual analysis of pupil contributions Categories.

CS: cross-sectional study; L: longitudinal study; OE: open-ended questions; MC: multiple choice; Int: intervention study; CL: curriculum linked study

APPENDIX B

Scoring Guides and Items Used in the Changes of State Test

ITEM 219

[Item]

Carmen wants to know what happens to the mass of something when matter changes from one form to another. She puts three ice cubes in a sealed bag and records the mass of ice in the bag to investigate this. Once the ice cubes have melted, she records the mass of the water in the bag. Which of the following best describes the result?

a. *The mass of the water in the bag will be less than the mass of the ice in the bag.*
b. *The mass of the water in the bag will be more than the mass of the ice cubes in the bag.*
c. *The mass of the water in the bag will be the same as the mass of the ice cubes in the bag.*

Describe your thinking. Provide an explanation for your answer.

[Coding guide]
A. Multiple-Choice

Response	*New Code*
Option C	2C
Option A/B	1A
NO Response	0

B. Open-ended

Description	*Sample Student Response*	*New Code*
Both same mass (or weight) and same material explicitly	*"(Students chose c) Mass will be the same because they are the same substance but in a different form."* *"(Students chose c) The mass of the matter will be the same as the ice cubes' mass because only its state of matter changed and not the molecules themselves."*	3
Identity of material but says nothing about mass (or weight) changes	N/A	–
Both different (larger or smaller) mass (or weight) and same material explicitly	*"(Students chose b) the water will have more mass because the water will have more mass if turned into liquid."* *"(Students chose a) I think the mass of the water in the bag will be less because the pressure changes. It's still the same amount of water, but the molecules move."*	2B

(*Continued*)

Open-ended
(Continued)

Description	*Sample Student Response*	*New Code*
Same mass (or weight) but not explicit about same material	*"(Students chose c) I think that the mass will stay the same because matter cannot be destroyed or created and if nothing can escape the bag it will stay the same in mass at least."* *"(Students chose c) Even if matter changes to another, the mass will still be the same because there are the same number of molecules in there."*	2C
Same mass but not same material	*"(Students chose c) It stays the same because you put an amount in there & just because the molecules change doesn't mean the amount has."*	2A
Different (larger or smaller) mass (or weight) but not explicit about same material	*"(Students chose b) Mass is space taken up. So I think it's b, since water fills up the bag but the ice cubes take less mass since it's bundled up into."* *"(Students chose a) Ice is solid and weighs more than liquid."*	1A
Neither identity of material nor anything about mass (or weight) changes, but uses *melting* in a relevant way	N/A	–
Any response that indicates the term melting incorrectly used/I don't know/off-topic response	*"I don't know why."* *"I guessed."*	0
Blank		0

ITEM 286

[Item]

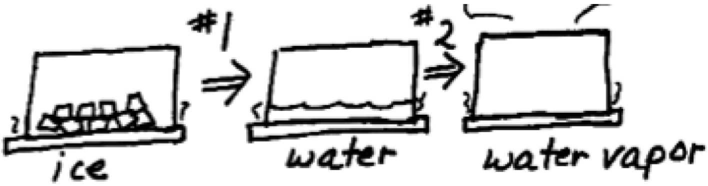
An iron block is melted into a liquid. What substance is the liquid?

[Coding guide]

Response	*New Code*
Response indicates it is the same material	2D
Response indicates it is a different material	1B
Response not related at all to melting of the iron	0

ITEM 214A

[Item]



Name each state of matter. ____________ __________ ___________

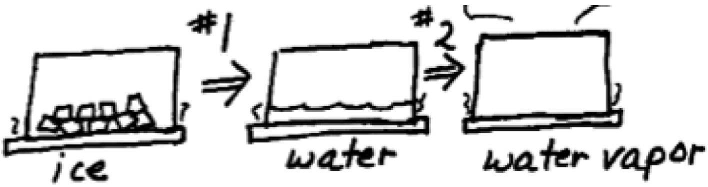
[Coding guide]
(for pictures 1 and 2)

Response	*New Code*
Response indicates correct use of solid and liquid	1D
Response indicates in-correct use of solid or liquid or fails to use terms	0
Blank	0

ITEM 215

[Item]

Explain what is happening to the molecules in phase change #1.

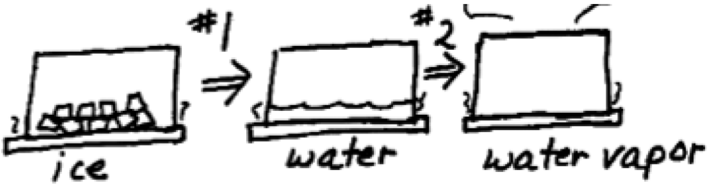


[Coding guide]

Responses	*Sample student response*	*New code*
Response indicates same material and (in some way) conservation/ continuity	*"The molecules are gaining energy and the ice is melting."* *"They are changing from still to moving."* *"The ice melted letting the molecules move around."* *"Water molecules are more spread out."* *"The molecules move very slowly as if frozen in place."* *"They absorb heat and loosen into more freely moving molecules in a liquid form"*	2D
Response indicates same material and (in some way) conservation/ continuity and says that weight doesn't change or volume decreases	N/A	N/A
Response talks of melting but no indication of conservation or continuity	*"The ice is melting into water."* *"The solids are melting and becomes a liquid after."* *"The ice melted."*	1D
Response talks about melting of molecules	*"Molecules melt."* *"The molecules are melting because solid ice cubes are transforming into a liquid water."* *"The molecules are starting to melt and change to the state of a liquid."*	0
Any response that indicates the term melting incorrectly used/I don't know/off-topic response	*"The molecules are crowded together."* *"The molecules join and create solid."* *"The molecules are moving even faster and evaporating the liquid"*	0
Blank		0

ITEM 214B

[Item]



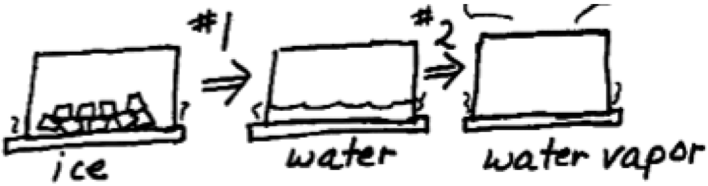
Name each state of matter. ____________ __________ ___________

[Coding guide]
(For pictures 2 and 3)

Responses	*New code*
Response indicates correct use of liquid and gas	1D
Response indicates incorrect use of liquid and gas or fails to use terms	0
Blank	0

ITEM 216

[Item]
Explain what is happening to the molecules in phase change #2.



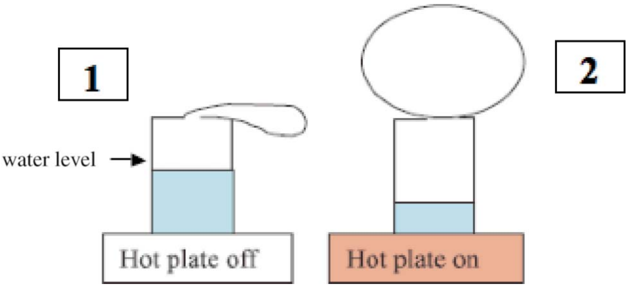
[Coding guide]

Description	*Sample Student Response*	*New Code*
Response indicates same material and (in some way) conservation/ continuity	*"The water is heated up and turning to gas form."* *"The water molecules take shape of the container and move freely."* *"The water absorbed heat energy which causes the molecules to move more quickly so it turns into gas."* *"The water is evaporating into the air and the molecules are slowly increasing in its speed while becoming less packed & flowing more."*	2D
Response indicates same material and (in some way) conservation/ continuity and says that weight doesn't change or volume decreases	N/A	–
Response talks of evaporation but no indication of conservation or continuity	*"The water is getting boiled and turns into water vapor."* *"The water evaporated into a gas."*	1D
Response talks about evaporation of molecules	*"The molecules are being vaporized."* *"The molecules are evaporating."*	0
Any response that indicates the term evaporation incorrectly used/I don't know/off-topic response	*"The molecules separate to form liquid."* *"The molecules are disappearing."*	0
Blank		0

ITEM 225

[Item]

A container with a little hole at the top is placed over a hot plate. There is water in the container. A deflated balloon is attached to the hole (1). The hot plate is turned on. The water starts boiling and the balloon inflates (2).



What is balloon (2) filled with?

a. *Air*
b. *Oxygen and hydrogen gas*
c. *Water vapor*
d. *Heat*

Explain your answer ____________________

[Coding guide]
A. Multiple-Choice

Response	New Code (Proposed)
Option C	2D
Option B	1B
Option D	1B
Option A	1B

B. Open-ended

Description	*Sample Student Response*	*New Code*
Balloon fills with water vapor, comes from water & is still water, so less liquid water	*"(Students chose c) The water decreases but it is just boiling into water vapor."* *"(Students chose c) Its water vapor b/c the heat turns the water into gas, and it fills the balloon and the water level is lower."*	2D
Balloon fills with water vapor coming from water, which is still water	*"(Students chose c) Since the water is evaporating the water turns into gas."* *"(Students chose c) The water was heated and it had so much energy it became water vapor and it fills the balloon."* *"(Students chose c)The water evaporates and the vapor blows up the balloon."*	2D
Balloon fills with gas that comes from the water but is no longer water	*"(Students chose a) The hot plate turns the water into gas & it releases the air into the balloon."* *"(Students chose b) The water turns into oxygen and hydrogen from evaporation."* *"(Students chose b) Water is H2O, so it becomes gas."*	1B
Mentions evaporation of water but does not relate to cause of balloon change		N/A
Balloon fills out for some other reason (e.g., air heated)	*"When the hot plate is on, it releases heat."* *"(Students chose b) The heat gives off oxygen so the balloon would inflate."* *"(Students chose d) The heat of the plate causes the balloon to expand."* *"(Students chose a) The balloon is filled with air once it inflates."*	0
Blank		0

ITEM 226

[Item]

Consider the combined mass of the container, water, balloon (deflated or inflated), and what the balloon contains. When the water boils,

a. *Mass stays the same because* ________________________________
b. *Mass decreases because* ____________________________________
c. *Mass increases because* ____________________________________
d. *There is no way to predict because* ___________________________

[Coding guide]

A. Multiple-Choice

Response	New Code
Option A	2C
Option B/C	1A
Option D	0

B. Open-ended

Description	*Sample Student Response*	*New Code*
Say mass stays the same and that it's still the same material.	*"Mass stays the same because the decreased water in Picture 2 actually is in the balloon in the form of water vapor."* *"Mass stays the same because the amount of molecules is the same."* *"Mass stays the same because the form changes but not the mass."* *"Mass stays the same because the mass is still the same as before, it just physically changed."*	3
Say mass stays the same, but don't specify whether it's still the same material	*"Mass stays the same because no air (water) really left the balloon."* *"Mass stays the same because no water leaked out of the balloon."* *"Mass stays the same because no one added anything."* *"Mass stays the same because nothing is added or subtracted.*	2C
Say it's still the same material, but mass decreases/increases	*"Mass increases because the balloon is filled with water vapor."* *"Mass decreases because water rises into the air."* *"Mass increases because it spreads, taking up more space."* *"Mass increases because the mass of water vapor goes up in the balloon."* *"Mass increases because the balloon inflates and makes the balloon bigger."* *"Mass increases because the water vapor is getting bigger."*	2B
Mention boiling/evaporate/or other relevant terms and say that it's a different material, no matter whether students think mass stays the same or not	*"Mass stays the same because water turns into gas."* *"Mass increases because gas molecules made the balloon expand."* *"Mass increases because the water makes gas."* *"Mass increases because the water boiling makes vapor."*	2A 1A

(*Continued*)

Open-ended
(Continued)

Description	*Sample Student Response*	*New Code*
Mention boiling/ evaporate/or other relevant terms Don't specify whether it's still the same material or not	*"Mass decreases because the hot plate boils up the water."* *"Mass decreases because of evaporation."* *"Mass decreases because the water evaporates."*	1A
No reference to boiling or irrelevant use of the term or response	*"Mass increases because object is put in the water."* *"Mass decreases because there's less water."* *"Mass increases because it [is] full of heat."* *"Mass decreases because some of the water is gone."* *"Mass increases because gas moves more freely."* *"There is not a way to predict because there's not enough information."* *"Mass increases because there's more air."*	1A 0
Blank		0