Using Progress Variables to Interpret Student Achievement and Progress
(BEAR Technical Report No. 2006-12-01)

Cathleen A. Kennedy & Mark Wilson
Berkeley Evaluation & Assessment Research (BEAR) Center
University of California, Berkeley

February 2007

We gratefully acknowledge the contributions of Nathaniel Brown, Karen Draney, Lydia Ou Liu, Diana J. Bernbaum, and Xiaohui Zheng. The research was part of a collaboration among the Berkeley Evaluation & Assessment Research (BEAR) Center, the Stanford Education Assessment Laboratory (SEAL), and the Center for Research on Evaluation, Standards, and Student Testing (CRESST) working under the auspices of the Center for the Assessment and Evaluation of Student Learning (CAESL). This material is based on work supported by the National Science Foundation under grant ESI-0119790 (CAESL). The findings and opinions expressed in this paper do not necessarily represent the views of the Foundation.

Abstract

Increasing demands for teacher responsibility for improving student performance have educators searching for new ways to diagnose student needs while learning is on-going. A promising approach described in this paper utilizes *progress variables* as the foundation of a coherent classroom environment that coordinates learning goals, instruction, and assessment. A progress variable is a representation of the knowledge, skills, and other competencies one wishes to increase through the learning activities associated with a curriculum (Wilson & Sloane, 2000). A framework for modeling student achievement and progress within a curricular unit using progress variables was developed to help teachers in day-to-day instructional planning . A process for establishing instructionally useful performance levels that provide an interpretive context for understanding student proficiency and instructional needs in a Science unit is described, and teachers' use of graphical representations of student proficiencies for instructional planning is illustrated.

Using Progress Variables to Interpret Student Achievement and Progress

Introduction

Increasing demands for teacher responsibility for improving student performance have educators searching for new ways to diagnose student needs while learning is on-going. A promising approach piloted in a study conducted by the Center for the Assessment and Evaluation of Student Learning (CAESL) utilizes *progress variables* (Masters, Adams & Wilson, 1990; Wilson, 2005) as the foundation of a coherent classroom environment that coordinates learning goals, instruction, and assessment. The project developed a framework for modeling achievement and progress within a curricular unit in Science to help teachers understand how the class as a whole is progressing toward curricular goals and to detect individual student needs. In this observational study we found that (1) learning goals, instruction, and assessment activities can be aligned through the use of progress variables, (2) instructionally useful performance levels can be established on a progress variable to provide an interpretive context for understanding student proficiency and instructional needs, (3) formative assessment activities can be designed to differentiate responses as indicators of particular performance levels, and (4) graphical representations of student proficiencies provide useful formative feedback to teachers for planning next steps in the classroom and for diagnosing individual student needs.

Progress variables, as implemented by the Berkeley Evaluation and Assessment Research (BEAR) Center (Wilson, 2005; Wilson & Sloane, 2000), are representations of the knowledge, skills, and other competencies one wishes to increase through the learning activities associated with a curriculum; they provide (a) the developmental structures underlying a metric for

measuring student achievement and growth, (b) a criterion-referenced context for diagnosing student needs, and (c) a common basis for the interpretation of student responses to assessment tasks. In the current study, progress variables were defined and then used to align assessment activities for a unit on buoyancy from the Foundational Approach to Science Teaching (FAST) Physical Science curriculum (Pottenger & Young, 1992), developed at the University of Hawaii at Manoa. The curriculum was selected because of its perspective of student development exemplified in its instructional approach (Pauls, Young & Lapitkova, 1999), its attention to teacher professional development (National Staff Development Council, 1999), and its strongly guided experiential approach that focuses on students constructing knowledge through hands-on activities (U. S. Department of Education, 2001). These hands-on instructional activities formed the basis for eliciting evidence of the progress variables from embedded assessment activities. Embedded assessment activities are assessment activities that are, to students, essentially indistinguishable from instructional activities, but which can generate immediate feedback to teachers and students about student performance, as well as reports for teachers, students, and parents detailing progress relative to expectations.

Two progress variables were developed to represent the progression of students toward important curricular goals in the unit. The first reflected student understanding of "Why Things Sink and Float" (the WTSF progress variable) and the second indicated the sophistication of reasoning students used in justifying their explanations of why things sink and float (the Reasoning progress variable). Each variable describes a continuum of student development from relatively naive understanding or ability to more sophisticated understanding, demarcated by several distinct performance levels that explain how understanding changes as students progress along the continuum from lower levels to higher levels.

When the study began, the unit included a multiple-choice pre/post test and a number of embedded assessment activities developed at the Stanford Education Assessment Laboratory (SEAL). The embedded assessment activities were used by teachers to informally ascertain student understanding through guided classroom discussions. One objective of the current project was to modify these assessments to provide more precise evidence and more interpretable feedback regarding the targeted progress variables. A number of the pre/post multiple choice items were modified to include written justifications of the selected responses, and the embedded assessment activities were modified to elicit evidence of student understanding of science concepts (as embodied in the progress variables) rather than of procedural skill.

This report begins by describing the development of the formative assessment system that was designed to improve teachers' understanding of how student knowledge develops in the unit. We describe how the interpretive context was integrated into the system to assist teachers in recognizing students' instructional needs while learning is on-going. We then give examples of how the system is used to diagnose student needs and discuss implications of this approach.

Method

Our first step was to develop progress variables and assessments using the BEAR Assessment System (BAS; Wilson & Sloane, 2000; Wilson, 2005) principles and building blocks. Progress variables were determined, items were designed to elicit evidence tied to specific levels of performance on the progress variables, progress guides were developed to associate potential responses on the items back to levels on the progress variables, and a measurement model was estimated to define the inferential structures needed to draw inferences about the progress variables from the student responses. The progress variables were calibrated to establish a consistent multidimensional scale to evaluate the development of proficiency over

time, and cut-points were set to differentiate performance levels on each progress variable to establish a quantitative interpretive context that could be used to map student development.

The assessment system was used in conjunction with the FAST buoyancy unit taught by eight teachers in 14 middle school science classes. From among their classes, each teacher selected a "target" class for which they would collect and report all student responses to the pretest, three Reflective Lessons (RL@4, RL@7, and RL@10), and the post test, for a total of 221 students. Four teachers also provided post test data for an additional 74 students, which were used for calibration of the progress variables but not to map growth. Student work from the five assessments was scored using the final versions of the progress guides described below (Figures 4 and 5). Scoring of the constructed response items was performed by a group of four raters familiar with the curriculum and trained in the use of the progress guides during a series of moderation sessions (Wilson & Sloane, 2000).

The BEAR Assessment System

The BEAR Assessment System is an integrated approach to developing assessments that provides meaningful interpretations of student work relative to the cognitive and developmental goals of a curriculum. It is grounded in four key principles guiding assessment development which are embodied in four building blocks (each associated with one of the principles; Wilson, 2005) that are tools for constructing a coherent system of meaningful curricular goals, instructional activities, and assessment. These principles and their associated building blocks are:

Principle 1: Assessment should be based on a developmental perspective of student learning.
Building Block 1: Progress Variables

Principle 2: What is taught and what is assessed must be clearly aligned.
Building Block 2: Items Design

Principle 3: Teachers are the managers and principal users of assessment data.
Building Block 3: Outcome Space & Progress Guides

Principle 4: Classroom assessment must uphold sound standards of validity and reliability.
Building Block 4: Measurement Model

These four principles also relate to the Assessment Triangle developed by the National Research Council Committee on the Foundations of Assessment and published in their report, *Knowing What Students Know* (National Research Council, 2001). The Assessment Triangle, shown in Figure 1, is a model of the essential connections and dependencies present in a coherent and useful assessment system. In this triangle, assessment activities (the *observation* vertex) must be aligned with the knowledge and cognitive processes (the *cognition* vertex) one wishes to affect through the instructional process, and the scoring and interpretation of student work (the *interpretation* vertex) must reflect measures of the same knowledge and cognitive processes. Meaningful connections among the three vertices – cognition, observation, and interpretation – are deemed essential for assessment to have a full impact on learning. The four building blocks of the BAS mentioned above (Progress Variables, Items Design, Outcome Space, and Measurement Model) map to the vertices of the Assessment Triangle and are also shown in Figure 1.

[Figure 1 here]

Earlier implementations of the BAS are described in Wilson and Sloane (2000) and Wilson and Scalise (2003). Those implementations involved developing an assessment system in conjunction with the development of a curriculum. The current study represents the situation one finds more commonly, where the curriculum predates the assessment system. In this case, we had to identify the developmental paths students were expected to follow based on the

instructional content, select and define relevant progress variables from among a range of options, and then adapt embedded formative assessment activities to align with those progress variables. This involved an iterative process in which the progress variables, item designs, outcome spaces, and measurement models were developed or chosen in tandem. Initial versions of these components were pilot tested with several classes using the FAST curriculum in the summer of 2003. Student responses from these pilot classes provided the data that guided the iterative revision of the assessment tasks before they were implemented in the regular school year.

Progress Variables

A progress variable is used to represent a cognitive theory of learning consistent with a developmental perspective. This building block is based on the principle that assessments are to be designed with a developmental view of student learning. This means that the underlying purpose of assessment is to determine *how students are progressing* from having less knowledge or expertise to having more in the domain of interest. Thus, progress variables are selected to reflect development of knowledge in a particular domain, to summarize that development in a way that is broadly unidimensional, and also to address the formative purpose of classroom assessment.

The selection of progress variables drew on the expertise of many professionals participating in the project, including science teachers, science education researchers, and psychometricians. We attempted to answer three questions as we identified the progress variables: (1) How many variables are needed, (2) Which variables are most appropriate for the assessment purpose, and (3) How many performance levels should be differentiated for the assessment purpose?

Typical of a middle school science curriculum, the FAST unit on buoyancy embodies numerous learning outcomes, including the development of content knowledge, process skills, and inquiry abilities. The selection of progress variables was motivated by several goals of the project, including the demonstration and use of (a) more than one variable; (b) variables dealing with different knowledge types (e.g., declarative, procedural, schematic, and/or strategic as described by Li and Shavelson, 2001); and (c) at least one variable that was not curriculum-specific. An unavoidable tension when choosing progress variables is the tradeoff between (1) coverage, which tends to drive the creation of multiple progress variables representing every possible curricular goal, and (2) usability, which limits the total number of progress variables that can be realistically learned and implemented by teachers and students. For this study we decided to develop two progress variables, one curriculum-specific, related directly to content knowledge, and the other a more universal inquiry skill that could be applied to other curricula dealing with other science content. Selecting two progress variables from a constellation of choices was a challenging decision but was made easier when we also considered the formative purpose of the assessments. We wanted to select progress variables that would be useful throughout the unit, which reduced the options somewhat. Determination of the performance levels on each progress variable was guided by both the instructional content and consideration of how teachers usually identify students who need help. A complete discussion of how each progress variable was selected is beyond the scope of this paper, but can be found in Kennedy and Wilson (in press).

Why Things Sink or Float

The first progress variable was developed to represent the trajectory of learning the content knowledge covered in the unit. The design of this progress variable was motivated

largely by the sequence of instructional lessons in the unit. These lessons, called *investigations* to emphasize the experiential approach to learning embodied by the curriculum, were explicitly designed to follow a developmental learning trajectory leading to an understanding of buoyancy from the perspective of relative density.[1] Figure 2 illustrates the sequence of twelve investigations in the unit, which appear along the base of the chart, and the instructional focus of each investigation which appear in the vertical columns. The figure lays out segments of the developmental learning trajectory intended by the curriculum developers.

[Figure 2 about here]

We found that the topic of "why things sink or float" was an important theme throughout the curriculum and therefore the most useful content area for charting student progress over the course of the entire unit. Once the progress variable domain was identified, qualitatively distinct performance levels were defined. The levels were initially guided by the curriculum trajectory, as shown in Figure 2. Review of student responses to early versions of the assessments suggested, however, that specifying performance levels developmentally below those that appear on the curriculum trajectory would be especially useful to teachers in diagnosing students' instructional needs. The levels "Has productive misconceptions about why things sink or float," "Has fundamental misconceptions about why things sink or float," and "Does not appear to understand any aspect of why things sink or float" were added particularly to address the formative assessment purpose. A map of the WTSF progress variable showing the performance levels is shown in Figure 3.

[Figure 3 about here]

[1] Note that there is an alternative approach, which uses the concept of "the buoyant force" as its central concept.

Reasoning

Items developed to elicit student understanding of WTSF (described in the *Items Design* section) prompted students to write justifications for why they thought a given item would sink or float. The existence of these written justifications prompted the development of a progress variable to represent a learning trajectory describing the increasing sophistication of reasoning displayed in those explanations. Although the curriculum does not specifically address the development of reasoning used in constructing explanations, the instructional content modeled the use of general principles to explain scientific phenomena. Discussions with project colleagues and examination of student explanations at different points in the unit led to the identification of six performance levels on this variable, from "Cannot formulate an explanation" at the lowest level to "Uses general principles" at the highest. The progress variable map is shown in Figure 4. This progress variable deals with a different kind of knowledge than science content knowledge and is intended to apply broadly to other curricula and assessments, wherever written justifications are required.

[Figure 4 about here]

Items Design

The items design building block is a framework for designing tasks to elicit specific kinds of evidence about student knowledge, as described in one or more progress variables. The guiding principle is that assessment should be seamlessly integrated into the instructional activities of a course. That is, assessment is not merely tacked on at the end of instructional units, but is embedded in normal classroom activity and may even be, from the student's point of view, indistinguishable from instruction (Black, Harrison, Lee, Marshall & Wiliam, 2002; Black, Harrison, Lee, Marshall & Wiliam, 2003; Black & Wiliam, 1998a; Black & Wiliam, 1998b).

At this stage we also answered three questions to formulate designs for the items, (1) What is the assessment purpose, (2) When should assessment take place, and (3) What types of assessment tasks are consistent with instruction and provide the kinds of evidence needed for the assessment purpose? Because the curriculum relied heavily on building upon prior knowledge, the primary purpose of the formative assessments was to determine students' readiness for progressing to the next part of instruction. Earlier research identified four critical junctures, so called "joints," where formative assessment data would be most useful. "The team came up with three criteria to identify the natural joints: (1) a subgoal of the end-of-unit goal is achieved, that is, there is a body of knowledge and skills sufficiently comprehensive to be assessed; (2) teachers need to know about student understanding before they can proceed with further instruction; and (3) feedback to students is critical to help them improve their understanding and skills of the material already taught" (Shavelson, SEAL & CRDG, 2005, p. 6). These junctures were after investigations 4, 6, 7 and 10. For the current study, the assessment after investigation 6 was not included in the analyses because it involved students generating a concept map, which could not be readily aligned with either progress variable.

Pretest

The existing 28-item multiple-choice pretest was used largely as is, with two modifications. The first modification was the inclusion of ten released items from the National Assessment of Educational Progress (NAEP) and Trends in International Mathematics and Science Study (TIMSS) assessments as a proxy for large-scale assessment. The second modification was to add written justifications, asking students to explain their choices, to the nine items that could be directly related to the WTSF progress variable. This was done primarily

to elicit more evidence of student conceptions about why things sink or float. Only the multiple-choice items with justification (MCwJ) were used in the current study.

Embedded Assessments (Reflective Lessons)

The original embedded assessments included three problem formats: (1) a graphing type of problem in which students create a data table and then draw a graph of the data, (2) multiple-choice items, and (3) performance items that assessed student skill in using scientific tools such as graduated cylinders and balances. Our analysis found that while the graphing and performance item types were good reflections of the day-to-day instructional activities, they did not provide much evidence of either progress variable. We introduced constructed-response activities into the new embedded assessment activities, which became known as Reflective Lessons, with two goals in mind: First, we hoped that both students and teachers would have more opportunities for reflection about what had been learned and what still needed to be learned while instruction was on-going. Second, we hoped that the additional information contained in the constructed responses would help teachers determine the source of student misunderstandings; teachers could then select activities that might be most useful to improve student learning in their particular classrooms.

We modified the graphing items by turning the focus toward using data to explain floating and sinking. Instead of assessing how well students record data and construct graphs, students were provided with a data table and a graph and asked to interpret them. The graphing items for each Reflective Lesson ask questions relating to recently covered concepts. For example, the assessment after Investigation 4 asks, “Explain in detail what the data and graph tell you about mass and depth of sinking.” The multiple-choice items were replaced by the open-ended essay question, “Explain why things sink and float.” This question is the same on each

Reflective Lesson. The performance activities were replaced by Predict-Observe-Explain (POE) activities. Originally conceived by Champagne, Klopfer and Anderson (1979) and further refined by Gunstone and White (1981), POE activities probe student understanding by asking students to predict the outcome of an experiment, describe what they see happen when the experiment is conducted, and finally explain any conflicts between their prediction and what they observed. For the current study, students were also asked to explain their predictions. A second POE was introduced into each Reflective Lesson to explore more advanced understanding. Students are asked to predict an event that depends on concepts to be taught in the next investigation.

Initial psychometric analyses revealed that the Reflective Lesson items, because they had been designed with the WTSF progress variable in mind, provided the best information about student understanding of why things sink or float, while the original multiple choice items (without justification) provided the least information (Kennedy, Brown, Draney & Wilson, 2005).

Post test

The original post test (i.e. the pre/post test) was replaced by a composite of multiple choice with justification items from the pretest and a selection of the Reflective Lesson items. We were not restricted to using the same items for the pretest and post test because we intended to calibrate both instruments onto the same multidimensional item response model (MIRM) scale.

At the conclusion of the items design, response categories for each item are aligned with the qualitative performance levels on the progress variables. These are essentially hypotheses about how the items are expected to generate responses; the hypotheses are subsequently tested with pilot data during the calibration process.

The Outcome Space and Progress Guides

The outcome space describes in detail the qualitatively different levels of item responses associated with a progress variable. This building block operationalizes the principle that teachers are to be the primary managers of assessment in the classroom. The purpose of the outcome space is to facilitate identification of student responses corresponding to a particular level on a progress variable. While a progress variable describes what students know and what they can do with that knowledge at several performance levels, the outcome space emphasizes the content of item responses that reveal those levels. The progress variable becomes the cognitive foundation, and the outcome space becomes the evidentiary foundation, for teachers to use on a daily basis in their classrooms on both formal assessments and in informal instructional contexts. Teachers' judgments about individual and group placement on the outcome space will influence many of their instructional decisions in the classroom at both the individual and group level.

Progress Guides are the tool teachers use to interpret student work. In some cases a progress guide is developed for each assessment task; in other cases a progress guide is used for all assessment tasks with annotated exemplars for individual tasks. In either case, each progress variable has its own progress guide(s). Taken together, the collection of progress guides for a particular progress variable maps the outcome space. The progress guides for the constructed response items in the FAST curriculum were initially developed by examining student responses to the Reflective Lessons collected during the pilot study. One progress guide was developed for each variable.

The attempt was made to associate each response with a single performance level of each of the progress variables, using only the evidence in the response to draw inferences about the

student's locations on the variables. This process initially revealed (a) responses that were consistent with the hypothesized progress variables, (b) responses that did not map to the progress variables, and (c) levels on the progress variables that were not observed in any of the responses. These findings resulted in revisions to early versions of the progress variables and progress guides which are described in Kennedy and Wilson (in press). The final versions of the WTSF and Reasoning progress guides are described here.

A progress guide contains much more detail, and possibly more levels, than the map of the progress variable (Figures 3 & 4) because the progress guide has to represent all possible student responses and must therefore deal with incomplete, incorrect, and unusual responses that might be observed. In particular, the WTSF progress guide (Figure 5) must deal with: (a) incorrect relationships, in which the correct concepts are used incorrectly (such as claiming that more massive objects are more likely to float); (b) imprecise responses, in which one could suspect that the student may have the correct concept but is expressing it using the wrong scientific terms (such as "heavy" instead of "massive", or "heft" instead of "density"); and (c) off-topic responses, in which the student is not responding directly to the question being asked. These incorrect relationships, in which correct concepts are used incorrectly, were thought to represent the first steps in learning to use these new concepts. A minus sign was introduced so that teachers could indicate a problem with a student's response. For example, a response that uses the concept of mass incorrectly, claiming that more mass would cause an object to float, could be labeled as M–. The advantage of this system is that it provides the teacher with a degree of freedom to value certain things while retaining the core meaning of the levels of the progress variable.

The WTSF progress guide shown in Figure 5 contains three columns. The first column describes the code value teachers use to designate the level of the response that a student produced,[2] the second column describes what a student at that level knows and provides an example of a typical response at that level, and the third column describes how the response or the student thinking needs to change to indicate performance at the next higher level. This progress guide includes more categories than the number of performance levels defined for the progress variable; we added categories for No Response and Unscorable.

Figure 5 about here

A challenge that is often faced in evaluating student work is how to handle responses that indicate more than one level. For this study, the team decided to invoke a "Harshness rule" for the WTSF progress guide. We felt these responses were indicative of an incomplete understanding of the higher level and that the higher level had not yet been truly reached; consequently, these responses were scored at the lower of the two levels. In other circumstances, one might adopt the alternative rule.

The progress guide for the Reasoning variable is shown in Figure 6. Once again, the progress guide contains more detail than the map of the progress variable (Figure 4), and adds the No Response and Unscorable levels. In contrast to the WTSF progress guide, a "Leniency rule" was invoked for this guide to deal with responses that exhibited characteristics of more than one level. Leniency was chosen in this case because it was felt that a lower level of Reasoning might sometimes be employed for communicative reasons rather than cognitive reasons. Moreover, while students may sometimes use scientific terms without understanding,

[2] Note that we avoid the use of numeric codes. In practice, we find that teachers tend to derive the wrong meaning from numeric codes, for example, averaging the values, rather than associating the code with what it means relative to student understanding and progress.

we felt that any evidence of a higher Reasoning level could be considered as valid evidence that the student was capable of performance at that level of Reasoning.

[Figure 6 about here]

At the conclusion of defining the outcome space, each potential response to an item is aligned with a performance level on the associated progress variables. At this point in the development process, these performance levels are still qualitatively defined; specific cut scores have not been determined. At this point in the design, we do not worry about the mechanics of how the codes will be interpreted for MIRM analyses; our focus is on helping teachers recognize what a particular response indicates about student knowledge on a progress variable.

Measurement Model

In the BAS, the primary goal of selecting a measurement model is to optimize the interpretive quality of assessments. In order to provide a strong criterion-referenced interpretation of student proficiency, we place a priori interpretational constraints on the model. First, we require (ideally) that the order of item difficulties is the same for all respondents and second, we require that the order of respondents is the same for any subset of items. To accomplish this, we use a polytomous extension of the Rasch model (Rasch, 1961, 1980) because of its positive qualities for developing graphical interpretive tools (Wilson, 2005).

Rasch-based modeling provides a convenient way to develop estimates of person proficiency and item difficulty using the same scale. This subsequently facilitates the interpretation of person proficiency estimates using criteria from the item content. An example of such a scale is shown in Figure 7. In this example the center vertical line represents a progress variable continuum, with Xs on the left side representing student locations (or proficiencies) and the item identifiers on the right side representing the proficiency required to have a .5 probability

of attaining a response at that level or higher. The difficulties of several possible response levels on items A, B, C, and D of Reflective Lesson 4, appear on the right side of the line. The location of the "A.MorV" item response level on the right side indicates the proficiency level where it becomes more likely to have a response at the "Mass or Volume" level or higher than at a lower level on item A. The item response locations are derived from the cumulative category counts.

The notion of distance on the map relates to the probability of responding in particular ways to the items. We can say, then, that a student at the location illustrated on the figure as Y has a greater than .5 probability of responding at the "Misconceptions" level on item A, and a smaller than .5 probability of responding at the "MorV" level on the item. This example illustrates why it is so important for the data to fit the model. Without a good fit such interpretations could not be made with reasonable assurances. This interpretive structure also allows us to describe student growth in the context of items. We see that while at time point Y the student was embracing a misconception about buoyancy, at time point Z he or she had developed more sophisticated knowledge and understood how the relationship of mass to volume affects buoyancy.[3]

[Figure 7 about here]

When dealing with multivariate proficiencies, we apply the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson & Wang, 1997). Our approach, then, is to fit the data to the model, rather than seeking a model to fit the data. Clearly, this makes the quality of the items design and the actual item development critical.

Calibration

In order to calibrate the items from all of the assessments onto the same multidimensional scale, five post test forms were developed for the study. Figure 8 shows on which post test form

[3] Note that Wright Maps may show either locations or progress. In this example, we combine the two.

each item from the pretest and the Reflective Lessons appears. For example, the pretest multiple-choice with justification responses appear on Forms A, B, C, and D; Reflective Lesson 4 part A appears on Forms C and D; and Reflective Lesson 10 part D does not appear on any post test form. That particular item was calibrated by anchoring parts A, B, and C to calibrated values from the post test calibration and then calibrating the final item in a subsequent calibration of the Reflective Lesson 10 instrument alone. The part C item of Reflective Lessons 4, 7 and 10 was the same item and was calibrated as a single item. Whenever identical items appeared on multiple post test forms, the item was calibrated as a single item. The ConstructMap (Kennedy, Wilson & Draney, 2006) software we used for calibration does this automatically, as illustrated in Figure 9. One limitation of the calibration study design is that proficiency estimates for Form E involve only 4 items for each progress variable.

Figure 8 about here

Figure 9 about here

To represent the theoretical model of what the assessments were intended to measure, a two-dimension partial credit model (PCM; Masters, 1982) was used. Once the model is selected, fit can be evaluated, and the assessments can be calibrated onto the progress variables. Early analyses of the data indicated that the Mass Only and Volume Only performance levels on the WTSF progress variable could not be differentiated well, so these two performance levels were combined (Kennedy, Brown, Draney & Wilson, 2005). Similarly, Unconventional Features could not be differentiated from Productive Misconceptions, so these categories were also combined for the current study. Raters continued to indicate codes for the nine performance levels available in the progress guide, but these were reduced to six categories for MIRM analysis (the No Response category was considered an indication of missing data for both

progress variables). Details of the calibration procedure can be found in Kennedy, Brown, Draney and Wilson (2005).

We examined the psychometric properties of the post test to determine whether the model upheld the basic assumptions of item response theory (IRT) modeling: that each subscale is unidimensional, that higher scores on an item are associated with higher overall ability estimates, and that the items within each subscale are conditionally independent. These findings are discussed more fully in Kennedy (2006), but we report here that the selected model does not violate these assumptions and the data fit the model sufficiently well for estimating proficiency.

Once the items were calibrated from the post test data, item parameters were anchored for the pretest and the Reflective Lessons so that proficiency estimates could be generated as needed.

Setting Performance Levels

The next step in the study was to establish cut-points between the performance levels on each progress variable. In this case, we used a simple quantitative method based on the Thurstonian thresholds of the item response categories. Thurstonian thresholds, which are computed for each step of an item, indicate the proficiency required to achieve a response at that level or above on the item 50% of the time. For example, a threshold of 1.25 logits for step 2 of an item means that a person with a proficiency of 1.25 logits is equally likely to provide a response that will be scored in category 2 or above or a response below category 2. The cut-point between two performance levels is the midpoint between the means of the Thurstonian thresholds of the two levels. Note, however, that there are no Thurstonian thresholds below the step 1 category, so there is no mean threshold value for the lowest category and a midpoint between the first two categories cannot be identified. For this study, we used the lower bound of

a 67% confidence interval around the second category's mean to define the cut-point between the first and second categories. An illustration of the results of determining cut-points in this way is shown in Figure 10.

Figure 10 about here

Once these logit values are established, the interpretive context is set for mapping student achievement and progress.

Findings

The *Methods* section focused on the mechanics of using the BAS to: establish progress variables and descriptive performance levels, design items to elicit evidence of the levels, construct evaluation procedures for aligning possible responses on the items to performance levels on the progress variables, and estimate a measurement model to maximize useful interpretation of the measures that are inferred from the student responses. It also described how the interpretive context is transformed from qualitative distinctions of performance levels to quantifiable logit ranges on continuous progress variables. In this section we address the question of how that process is beneficial to teaching and to further improvement of the assessment design.

Learning goals, instruction, and assessment activities were aligned through the use of progress variables. One purpose of the study was to demonstrate the extent to which progress variables could be used to coordinate assessment with instruction and curricular goals. While the CAESL/FAST study is not a full implementation of the BEAR Assessment System, it is illustrative of how assessments are developed in the real world of limited resources and trade-offs. Due to a relatively short project timeline, a high percentage of the summative assessment

items were used as they were initially developed; that is, without consideration of the progress variables to be measured. Thus, these items provided less focused assessment evidence than would likely have been available had the items been developed in concert with the progress variables. We were, however, able to adapt/develop embedded assessment activities that were more closely coordinated with the progress variables.

Figure 11 is an illustration of how learning goals in the form of statements about what students know, instruction in the form of investigation topics, and occasions for assessment are aligned with the WTSF progress variable. The Reasoning progress variable was not developed to align as readily with instruction, although student progress over time was still anticipated.

Figure 11 about here

The progress variables were developed to reflect the most important learning goals of the curriculum and to define qualitatively distinct steps along the learning trajectories expected by the curriculum developers. We then developed embedded assessment activities containing items that were comparatively good measures of the progress variables. In addition, we added constructed-response segments to those original summative items that appeared to target the progress variables fairly well. We were also able to construct a post-test from a combination of the summative assessment items with their justification extensions and the new Reflective Lesson items. Thus, in a period of about a year, we were able to formalize the curricular goals into two progress variables with well-defined levels, modify and expand a collection of pre-existing items to more closely coordinate with the progress variables, define progress guides to evaluate student work, and apply appropriate measurement models for the purposes of examining student growth.

Instructionally useful performance levels are established for each progress variable to provide an interpretive context for understanding student proficiency and instructional needs. In addition to defining qualitatively distinguishable performance levels on each progress variable, we were able to determine quantitative cut-scores to differentiate the levels when reporting student progress. We were able to do this in a relatively short amount of time by using Thurstonian thresholds. In most cases, we prefer to follow a more subjective approach that relies on the judgments of teachers, curriculum developers, and educational researchers but we did not have sufficient time to implement that approach here. Performance levels represent curriculum-specific standards based on learning objectives and these depend on value judgments and expectations. We are following this method, which we refer to as Criterion Mapping, with sixteen science modules from the Full Option Science System (FOSS) developed at Lawrence Hall of Science. These modules are currently deployed in dozens of classrooms across the country and we are gathering data to evaluate this method.

Nevertheless, the quantitative approach used in the current study did produce realistic performance levels that were used to map student achievement and progress. An example of a progress map on the WTSF variable for a class in the study is shown in Figure 12. The map displays the names of the instruments administered at each time point along the x-axis at the top of the chart. Each individual point on the map shows the average of the student proficiency estimates for the class and indicates the concepts that students were working to understand at a particular time point. The horizontal shaded bands that run across the map represent the performance levels on the progress variables for the curriculum.

Figure 12 about here

Initially, students were primarily using misconceptions or nuances about mass to explain sinking and floating, and by the end of the unit students were using the density of objects to explain sinking and floating. The class in general did not quite attain the level of performance anticipated by the curriculum designers, which was to use relative density to explain why things sink and float.

Figure 13 shows progress on the Reasoning progress variable for the same class. In this case, we find that students in the class corrected their approach to constructing explanations early in the course. Although they initially tended to use inadequate explanations, by investigation 7 they were using general principles to explain their answers. The drop-off in the students' performance in the post test is not unexpected, as there is less scaffolding for these items compared to the items in the Reflective Lessons.

Figure 13 about here

Improving the alignment of items to particular performance levels and obtaining input from curricular experts would probably improve the accuracy of the cut-points, but we were able to demonstrate how an interpretive context is developed and then applied to actual data gathered from students as they engaged in the assessment activities in this unit.

Formative assessment activities can be designed to differentiate responses as indicators of particular performance levels. The use of progress variables and performance levels also provides a context for evaluating the contribution of individual items to the assessment purpose. For example, the inherent purpose of assessments administered after particular investigations is to ascertain whether student understanding is near that level. Items on RL4 should help differentiate students who are using mass to explain floating and sinking from those who are still working with misconceptions. Similarly, items on RL7 should help

differentiate students who understand the relationship of mass to volume from those who do not, and items on RL10 should differentiate students who understand the effect of density on the buoyancy of objects placed in water from those who are still using mass and/or volume to explain floating and sinking.

When we examine the Thurstonian thresholds of the items from the RL4 assessment, shown in Figure 14, we find that all of the items differentiate between the "Misconceptions" and "Mass" response categories sufficiently well. In addition, all of the items differentiate responses at the "Mass" level from responses as the "Mass & Volume" level. We find, however, that a response at the "Mass" level on item RL4D requires quite a lot more proficiency than providing a response at that level on items RL4A, RL4B and RL4C. There are a couple of likely explanations for this. We know that the item deals with concepts that have not yet been introduced, but the structure of the item may not even permit a response using mass in the explanation. In fact, item RL4D asks students to predict whether an object will have a smaller or larger depth of sinking when the same mass is packaged into straws of different volumes. The item states, "We know that 2 g of BBs would make our original straw depth of sinking about 8 cm. Now, what is going to happen when we put these 2 g of BBs into a fat straw? Predict the depth of sinking of the fat straw using the same 2 g of BBs. How do you know?" When students know that the mass is the same for the two situations, they may be less inclined to use mass to explain their answers. We might conclude from this that item RL4D is better at differentiating knowledge of how volume affects buoyancy than at differentiating knowledge of how mass affects buoyancy.

Figure 14 about here

When we look at the graph of Thurstonian thresholds for the Reasoning progress variable, shown in Figure 15, we find that the items only differentiate the highest performance level, and item RL4C does not really differentiate levels of Reasoning well at all.

Figure 15 about here

Graphical representations of student proficiencies provide useful formative feedback to teachers for planning next steps in the classroom and for individual diagnosis. The model of progress variables with quantitatively defined performance levels allows reporting of MIRM proficiency estimates in an interpretable format. Rather than providing teachers with numeric proficiency estimates, or even total test scores, the reports generated in this system indicate a range of concepts and skills that are within the reach of student understanding at a particular point in time. For example: the *Frequency Map* (Figure 16), which is essentially the left (respondent) side of a Wright Map, shows proficiency estimates at a given time for a whole class, while the *Performance Map* (Figure 17) shows proficiency estimates over time.

A teacher might use *Frequency Maps* most often to get a general sense of how students in a class are performing relative to instructional goals. Figure 16 is an example of a *Frequency Map* on the WTSF variable for one class after Reflective Lesson 7. Several students are not operating at the expected level for this point in the unit, but the students demonstrating the lowest level of understanding are of the most concern to the teacher. When these maps indicate that some students are falling behind, the teacher can request an *Ability Estimates by Level* report, which lists the names of students in each performance level for each variable. Once the teacher has identified students that may need additional support, he or she may request *Performance Maps* for those students to get an overview of performance over the duration of the curriculum thus far.

Figure 16 about here

Figure 17 shows the *Performance Map* for the WTSF progress variable for Amy after she has completed Reflective Lesson 7. We see that she was progressing as expected from the pretest to Reflective Lesson 4, but unexpectedly dropped back at Reflective Lesson 7. We see in Figure 18, however, that her performance on the Reasoning progress variable did not show this pattern. Instead, her performance in Reasoning improved at each time point. Examining student performance on each variable allows teachers to develop a more complete picture of a student's strengths and weaknesses.

Figure 17 about here

Figure 18 about here

Some cases will warrant further examination of the student's performance at the individual item level to more completely understand the nature of a problem the student is having. To understand more about what happened with this student between investigations 4 and 7, the teacher can request a *Wright Map* for one student.

Figure 19 shows the student's performance on Reflective Lesson 7. The student's overall proficiency level is indicated by the X on the left side of the chart, while the proficiency required to respond at each level on items A, B, C and D are indicated on the right side of the chart. Item responses that appear near to the student's proficiency level are those the student is expected to achieve about 50% of the time (and *not* achieve about 50% of the time). We would expect this student to generate responses that use mass or volume alone to explain sinking and floating on items A, C and D about 50% of the time. Item responses that appear higher on the chart than the student's location are those that require more proficiency that the student is currently exhibiting. Thus, we would not expect this student to generate a response that uses the relationship of mass

to volume as an explanation for sinking and floating on those items. Because distance is interpretable on the logit scale, a teacher can see at a glance how far a student's current understanding is from some targeted level. This student appears to have moved beyond misconceptions to an elementary understanding of how mass or volume alone affects sinking and floating, and is not yet using the relationship of mass to volume. A review of the student's actual responses compared to these expected responses can then provide additional information about the specific needs of this student (for further discussion of the diagnostic maps produced by the ConstructMap software see Kennedy & Draney, 2006; Kennedy & Wilson, in press; and Kennedy, Wilson & Draney, 2006)

Figure 19 about here

Using reports such as these, teachers are able to see at a glance not only how a student is performing at a particular time, but also any trends over time. They are also able to see if there are strengths and weaknesses in a student's performance by looking at the maps for the individual progress variables. Unexpected variations in proficiency or other problems can be noticed early so that corrective steps may be taken while instruction is on-going. In addition, when students see their own progress represented visually, particularly relative to the performance levels, they may be better able to discuss their progress with their teachers or to take other steps to improve their performance (e.g., Roberts & Sipusic, 1999).

The multidimensional aspect of these maps also provides some advantages over unidimensional modeling. Educationally, the teacher and student now have multiple dimensions that can be interpreted (Wilson & Sloane, 2000). Statistically, there is the potential for increased precision of measurement (Adams, Wilson & Wang, 1997). And, in situations where proficiency estimates are based on a relatively small number of observations or items (as educational

variables often are), the use of the collateral information available in correlated variables can increase the reliability with which person proficiency is estimated (Wang, Wilson & Adams, 1998).

Conclusion

The approach of developing progress variables to represent the central learning goals of a curriculum, and then using the progress variables to guide the development of embedded assessment activities, appears to be useful for the design of a coherent classroom assessment system. Progress variables provide a common basis for interpreting performance on different tests and examining progress over time. Embedded assessment activities targeting assessment of particular performance levels on the progress variables facilitate teachers' monitoring of student progress and their ability to compare the current state of learning with the expectations of the curriculum. This approach provides a system that helps teachers make educationally-important decisions, including the identification of essential concepts that have not been learned well enough by most students to support their progress to the next phase of instruction.

A current study is using this technique in 30 classrooms in Fall 2006 to obtain teacher perceptions about the feasibility and usefulness of the approach in day-to-day classroom practice (Assessing Science Knowledge).

References

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

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom*. London, UK: King's College London Department of Education and Professional Studies.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.

Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Black, P. & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-148.

Champagne, A. B., Klopfer, L. E. & Anderson, J. (1979). *Factors influencing the learning of classical mechanics*. University of Pittsburgh.

Gunstone, R. F., & White, R. T. (1981). Understanding of Gravity. *Science Education, 65(3)*, 291-299.

Kennedy, C. A. (2006). Simplified Scoring of Performance Activities: Comparing Assessment Stories from Complex and Simple Scoring Approaches, *National Council on Measurement in Education Annual Meeting*. San Francisco, CA.

Kennedy, C. A., Brown, N. J. S., Draney, K. & Wilson, M. (2005). Using progress variables and embedded assessment to improve teaching and learning, *American Educational Research Association Annual Meeting*. Montreal, Canada.

Kennedy, C.A. & Draney, K. (2006). Interpreting and using multidimensional performance data to improve learning. X. Liu and W. Boone (Eds.) *Applications of Rasch Measurement to Science Education*. Chicago: JAM Press.

Kennedy, C.A. & Wilson, M. (in press). Using progress variables to map intellectual development. In R. Lissitz (Ed.) *Proceedings of the Assessing and Modeling Cognitive Development in School: Intellectual Growth Standard Setting Conference*. College Park, MD. October 19-20.

Kennedy, C. A., Wilson, M. & Draney, K. (2006). *ConstructMap v4.3*. [computer program] University of California, Berkeley Evaluation & Assessment Research Center.

Li, M. & Shavelson, R. J. (2001). Examining the links between science achievement and assessment. Paper presented at the AERA Annual Meeting, Seattle, WA.

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

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

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

Mead, R. (1976). *Assessing the fit of data to the Rasch model*. Annual Meeting of the American Educational Research Association. San Francisco, California.

National Research Council (2001). *Knowing what students know*. Committee on the foundations of assessment. J.W. Pellegrino, N. Chudowsky, R. Glaser (Eds.) Washington, D.C.: National Academy Press.

National Staff Development Council (1999). *What works in the middle: Results-based staff development*. Retrieved 11/14/06 from http://www.nsdc.org/midbook/foundation.pdf.

Pauls, J., Young, B., Donald, & Lapitková, V. (1999). Laboratory for Learning. *The Science Teacher, 66 (1)*, 27-29.

Pottenger, F. & Young, D. (1992). *The local environment: FAST 1 Foundational Approaches in Science Teaching*. University of Hawaii Manoa: Curriculum Research and Development Group.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. *4*, 321-334.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago. University of Chicago Press (original work published 1960).

Roberts, L., & Sipusic, M. (Writer), & M. Sipusic (Director) (1999). *Moderation in all things: A class act [VHS]*. In L. Roberts (Producer). Berkeley, CA University of California, Berkeley Evaluation & Assessment Research Center.

Shavelson, R, Stanford Educational Assessment Laboratory (SEAL) and Curriculum Research & Development Group (CRDG). (2005). *Embedding Assessments in the FAST Curriculum: The Romance between Curriculum and Assessment*. Final Report. Palo Alto, CA: Stanford University.

U. S. Department of Education Expert Panel on Mathematics and Science Education (2001). Retrieved 11/15/06 from http://www.ed.gov/offices/OERI/ORAD/KAD/expert_panel/newscience_progs.html

Wang, W., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 139-155). Norwood, NJ: Ablex Publishing.

Wang, W., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with multidimensional Rasch models. *Journal of Outcome Measurement*, *2*, 240-265.

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

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

Wilson, M. & Scalise, K. (2003). Reporting progress to parents and others: Beyond grades. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom*. NSTA Press: Arlington, VA, 89-108.

Figure 1. The National Research Council's Assessment Triangle with associated BEAR Assessment System building blocks.

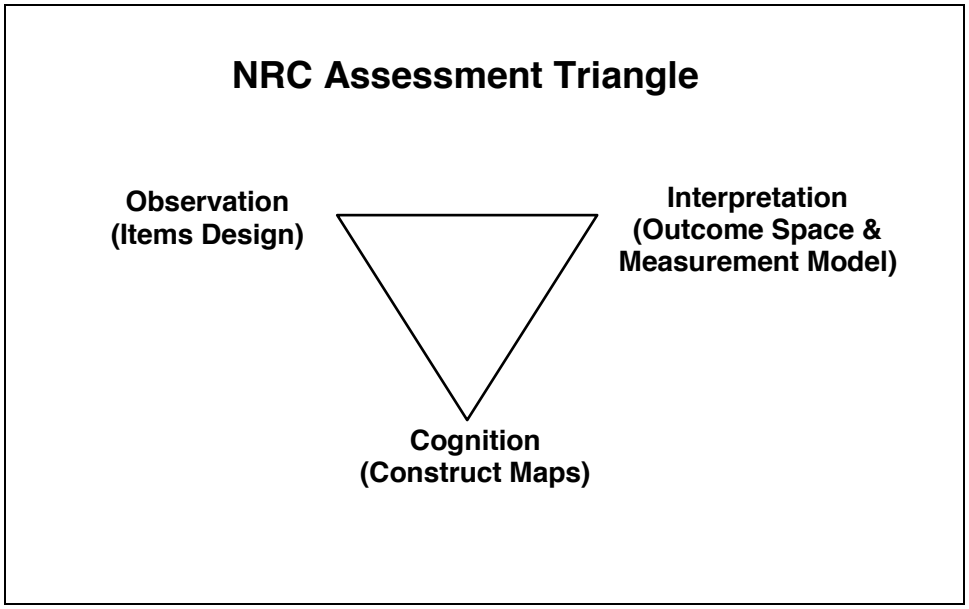

Figure 2. Sequence of twelve investigations in the Buoyancy unit and the associated developmental learning trajectory.

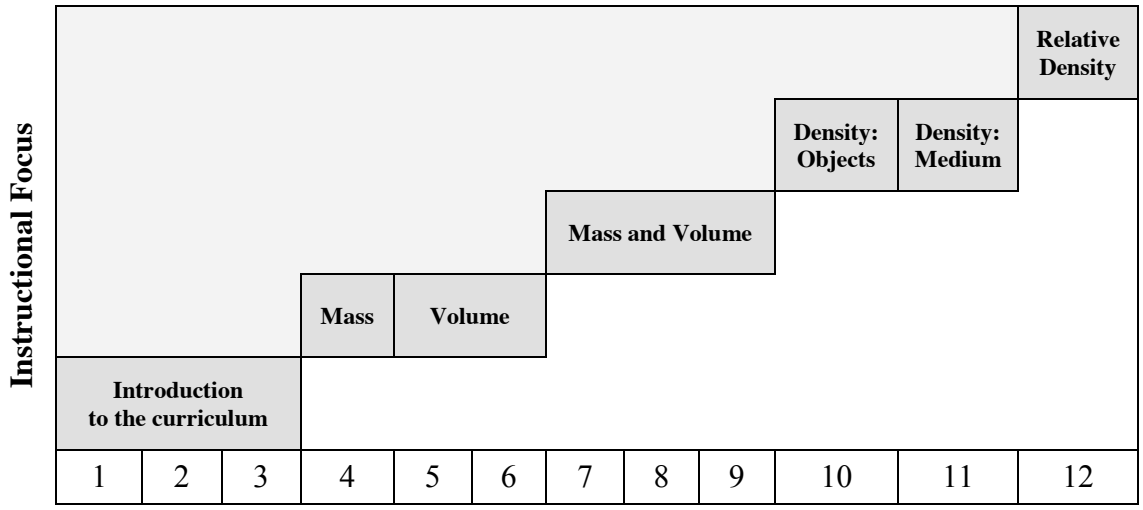

Figure 3. Map of the WTSF progress variable with qualitative performance levels.

What the student knows about Why Things Sink and Float
Knows how relative density affects floating and sinking in different liquids.
Knows how density affects floating and sinking in water.
Knows how the relationship of mass to volume affects floating and sinking.
Knows how volume affects floating and sinking when mass is held constant.
Knows how mass affects floating and sinking when volume is held constant.
Has productive misconceptions about why things sink or float.
Has fundamental misconceptions about why things sink or float.
Does not appear to understand any aspect of why things sink or float.

Figure 4. Map of the Reasoning progress variable with qualitative performance levels.

The kind of reasoning the student uses in constructing explanations
Knows how to use an explicit principle that applies to objects in general to explain an answer.
Knows how to use a specific relationship in which the object, the property, and the magnitude of the property (e.g., more vs. less) are all clear.
Knows how to use a specific relationship, but the object, the property, or the magnitude of the property (e.g., more vs. less) is not clear.
Knows how to use prior experience, in the form of a personal observation or an authoritative source, to explain an answer.
Cannot formulate an adequate explanation, but instead either restates their answer as an explanation, or simply asserts that their answer is correct.
Cannot formulate an explanation for their answer.

Figure 5. Final version of the progress guide for the WTSF progress variable. Example responses are hypothetical responses to the WTSF essay used for illustration. Actual responses were generally much longer.

Harshness Rule:
If different parts of the response suggest different levels, score the **lowest possible level**.

Level	What the Student Already Knows	What the Student Needs to Learn
RD	**Relative Density** Student knows that floating depends on having less density than the medium. • "An object floats when its density is less than the density of the medium."	
D	**Density** Student knows that floating depends on having a small density. • "An object floats when its density is small."	To progress to the next level, student needs to recognize that the medium plays an equally important role in determining if an object will sink or float.
MV	**Mass and Volume** Student knows that floating depends on having a small mass and a large volume. • "An object floats when its mass is small and its volume is large."	To progress to the next level, student needs to understand the concept of density as a way of combining mass and volume into a single property.
M / **V**	**Mass** Student knows that floating depends on having a small mass. • "An object floats when its mass is small." **Volume** Student knows that floating depends on having a large volume. • "An object floats when its volume is large."	To progress to the next level, student needs to recognize that changing EITHER mass OR volume will affect whether an object sinks or floats.
PM	**Productive Misconception** Student thinks that floating depends on having a small size, heft, or amount, or that it depends on being made out of a particular material. • "An object floats when it is small."	To progress to the next level, student needs to refine their ideas into equivalent statements about mass, volume, or density. For example, a small object has a small mass.
UF	**Unconventional Feature** Student thinks that floating depends on being flat, hollow, filled with air, or having holes. • "An object floats when it has air inside it."	To progress to the next level, student needs to refine their ideas into equivalent statements about size or heft. For example, a hollow object has a small heft.
OT	**Off Target** Student does not attend to any property or feature to explain floating. • "I have no idea."	To progress to the next level, student needs to focus on some property or feature of the object in order to explain why it sinks or floats.
NR	**No Response** Student left the response blank.	To progress to the next level, student needs to respond to the question.

X	**Unscorable** Student gave a response, but it cannot be interpreted for scoring.	

Figure 6. Final version of the progress guide for the Reasoning progress variable. Example responses are hypothetical responses to the WTSF essay used for illustration. Actual responses were generally much longer.

Leniency Rule:
If different parts of the response suggest different levels, score the **highest possible level**.

Level	What the Student Can Already Do	What the Student Needs to Improve
P	**Principled** Student uses an explicit principle that applies to objects in general. • "An object floats when its mass is large."	
R	**Relational** Student uses a specific relationship in which the object, the property, and the magnitude of the property (e.g., more vs. less) are all clear. **Note:** Some of the parts of the relationship may be made clear by the item stem, or by another part of the response (e.g., a prediction), rather than in the explanation. • "Object A floats because its mass is large."	To progress to the next level, student needs to use a principle that would apply to objects in general.
U	**Unclear Relational** Student uses a specific relationship in which either the object, the property, or the magnitude of the property (e.g., more vs. less) is not clear. • "Object A floats because of its mass."	To progress to the next level, student needs to explicitly identify all three parts of the relationship in their explanation.
E	**Experiential** Student justifies their answer by appealing to prior experience, in the form of a personal observation or an authoritative source. • "It floats because that's what we saw in class."	To progress to the next level, student needs to use a relationship to explain their answer, not just evidence to justify it.
IE	**Inadequate Explanation** Student either restates their answer as an explanation, or simply asserts that their answer is correct. • "Object A will float."	To progress to the next level, student needs to understand what evidence is and the relationship between evidence and an explanation.
OT	**Off Target** Student cannot or does not give an explanation for their answer. • "I have no idea."	To progress to the next level, student needs to justify their answer in some way.
NR	**No Response** Student left the response blank.	To progress to the next level, student needs to respond to the question.

X	**Unscorable** Student gave a response, but it cannot be interpreted for scoring.	

Figure 7. An excerpt from a Wright Map showing a student's location at two time points, Y and Z, relative to item response categories on items from Reflective Lesson 4.

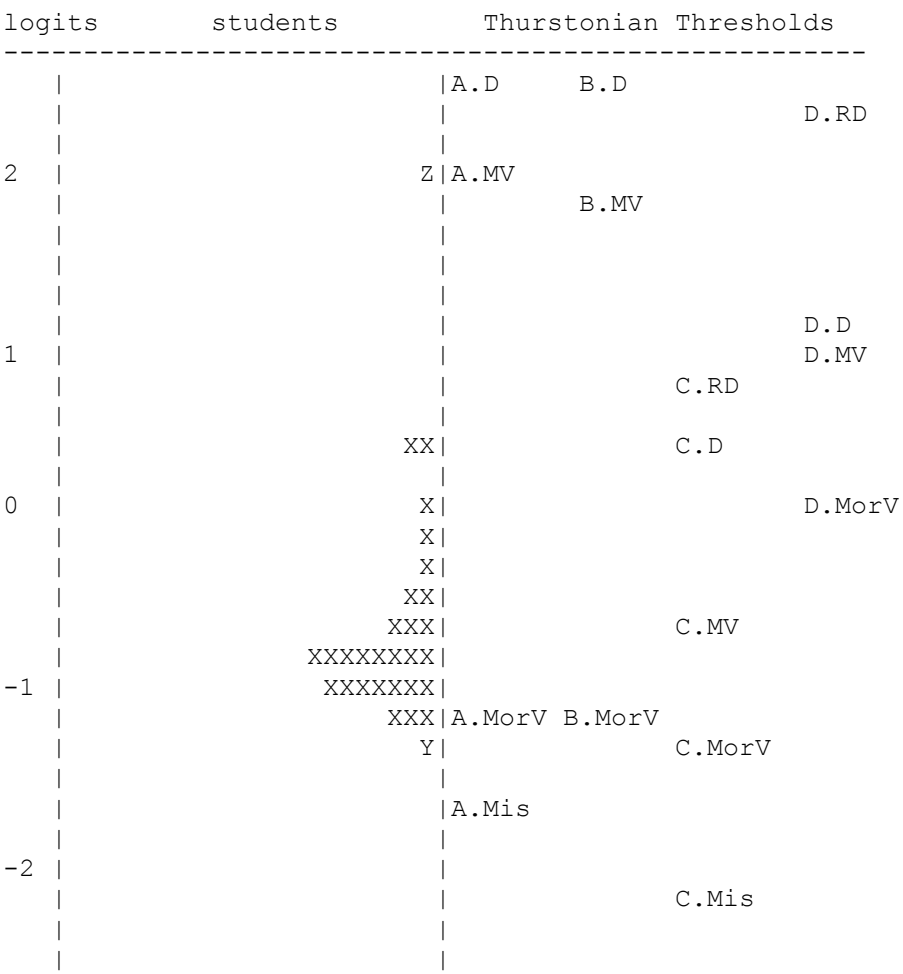

Figure 8. Distribution of pretest and reflective lesson items across five forms of the post test. The numeric value in the Pretest column indicates the number of multiple choice with justification items included on the post test. The shaded areas indicate the form an item was part of. RL4 part C, RL7 part C and RL10 part C were identical items and were calibrated as a single item.

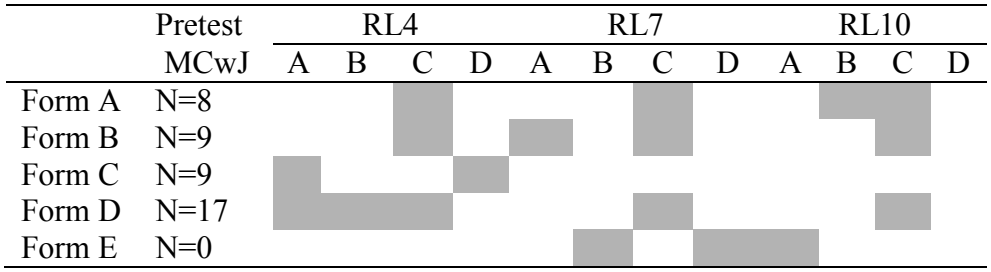

Figure 9. An example of how ConstructMap automatically aligns common items on multiple instruments. Part A of Reflective Lessons 7 and 10 include 2 items each.

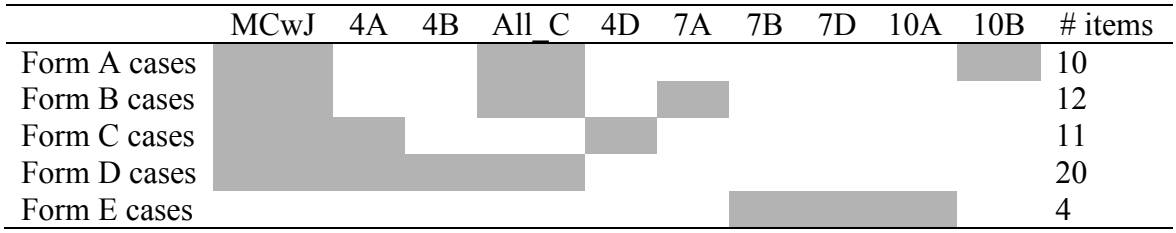

Figure 10. Means of the Thurstonian thresholds and cut-points for the performance levels (criterion zones) for the Buoyancy: WTSF progress variable.

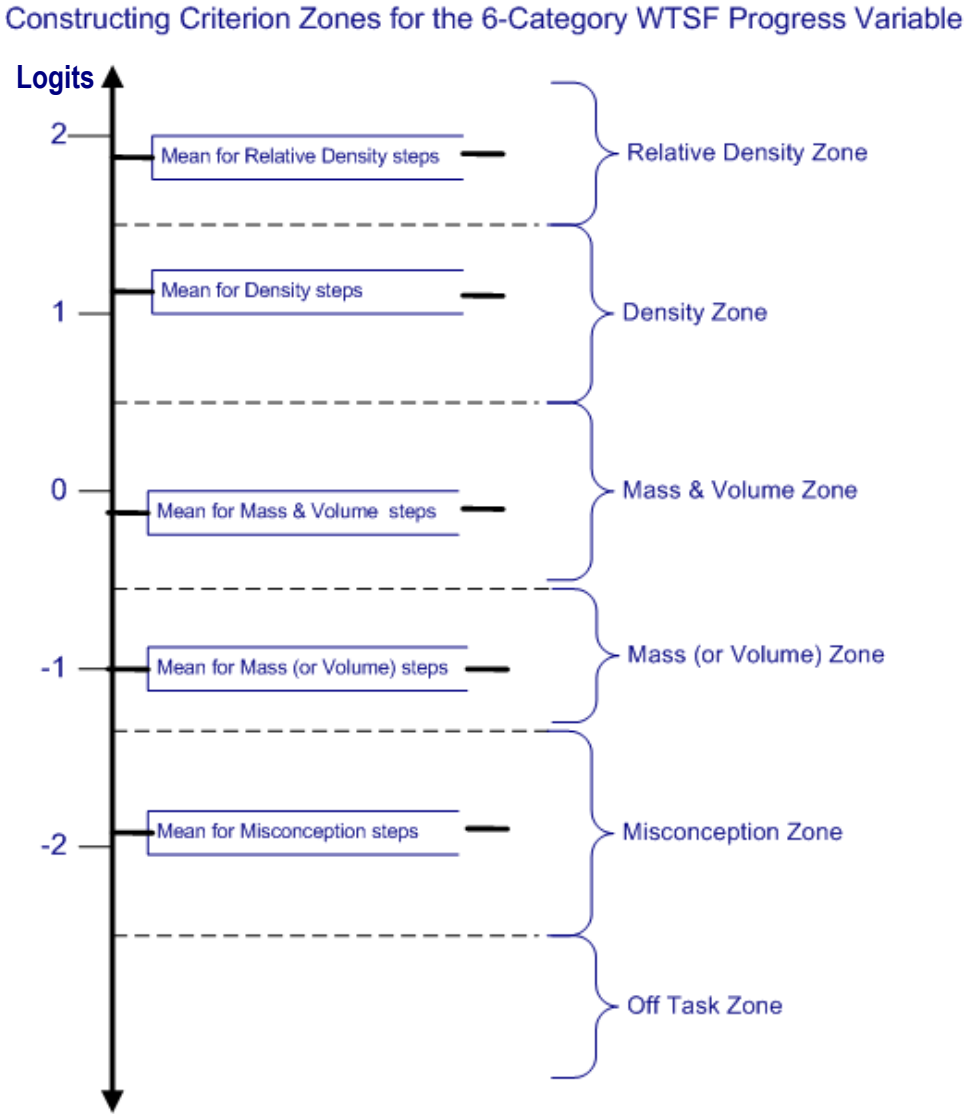

Figure 11. Alignment of learning goals, instruction and assessment using a progress variable.

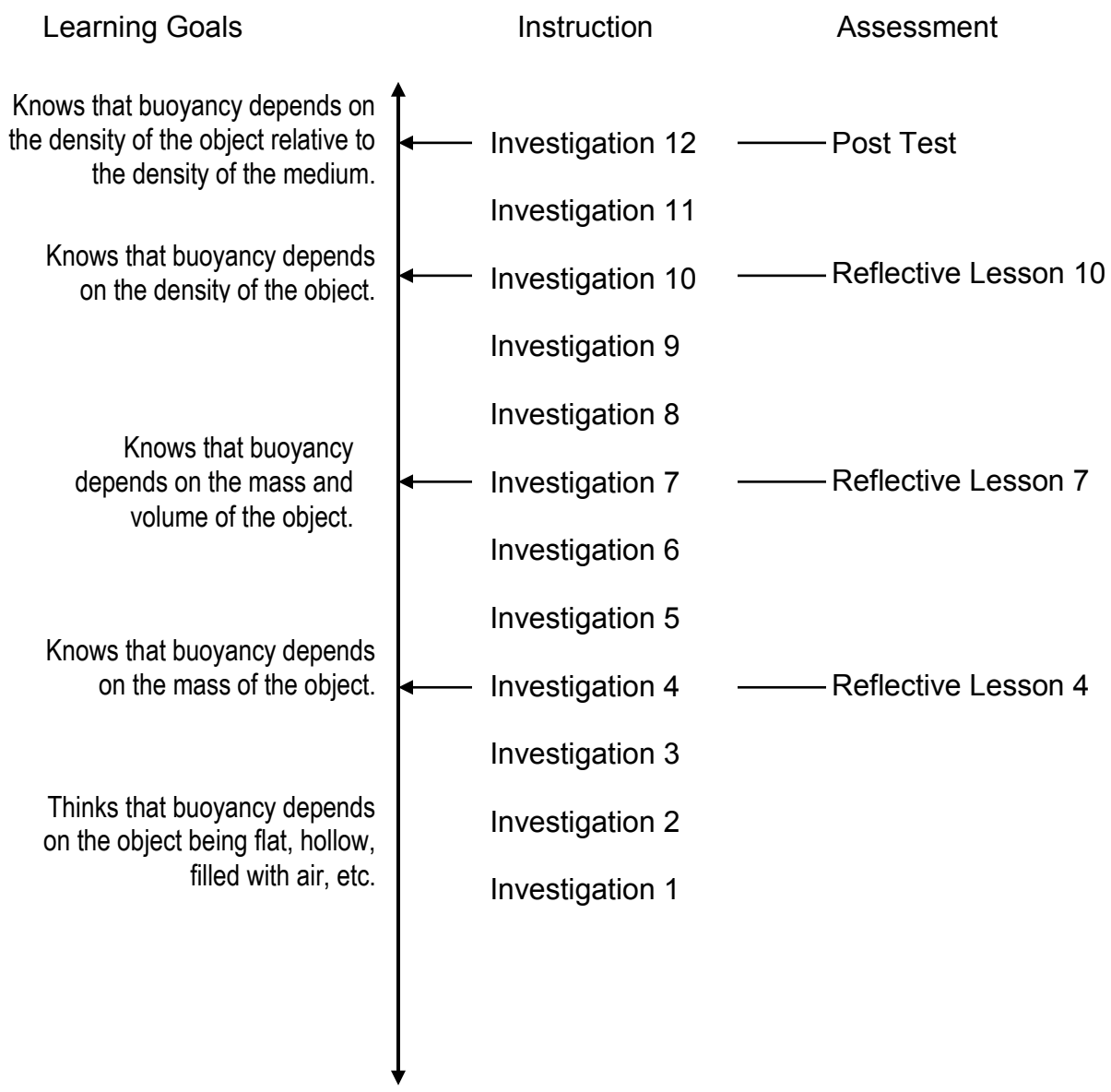

Figure 12. Map of average student progress on the WTSF progress variable for the students of Teacher 3.

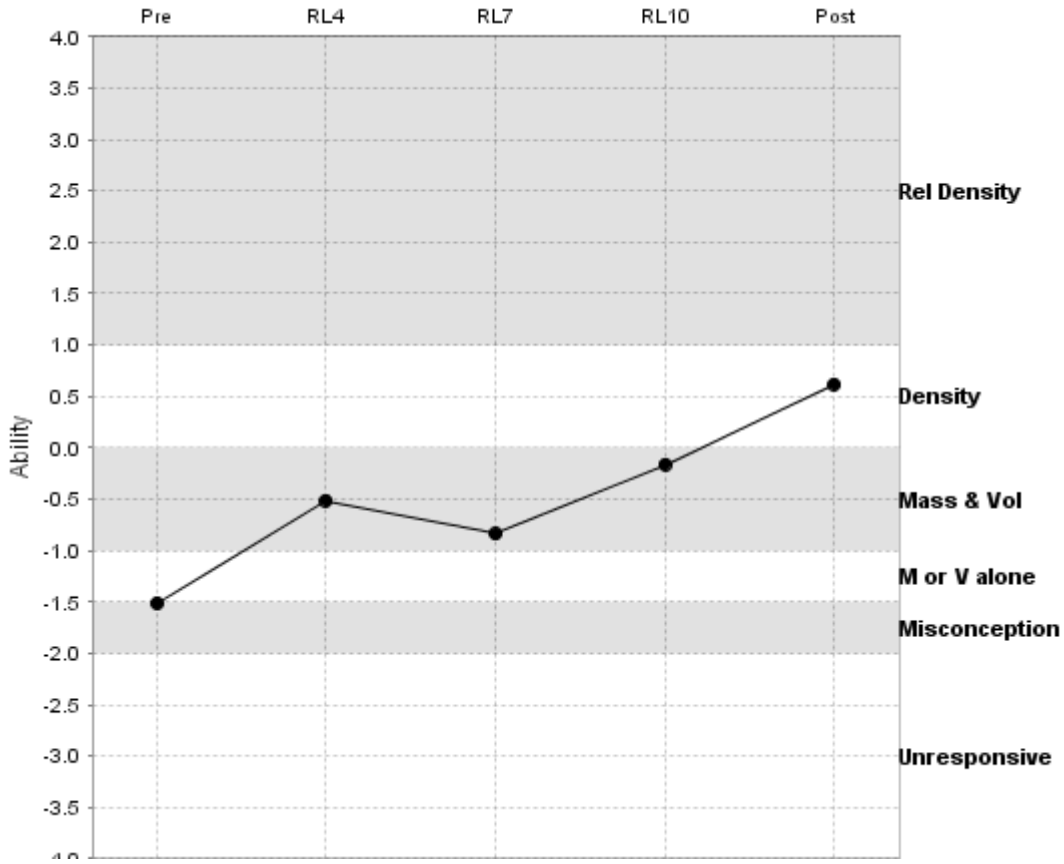

Figure 13. Map of average student progress on the Reasoning progress variable for the students of Teacher 3.

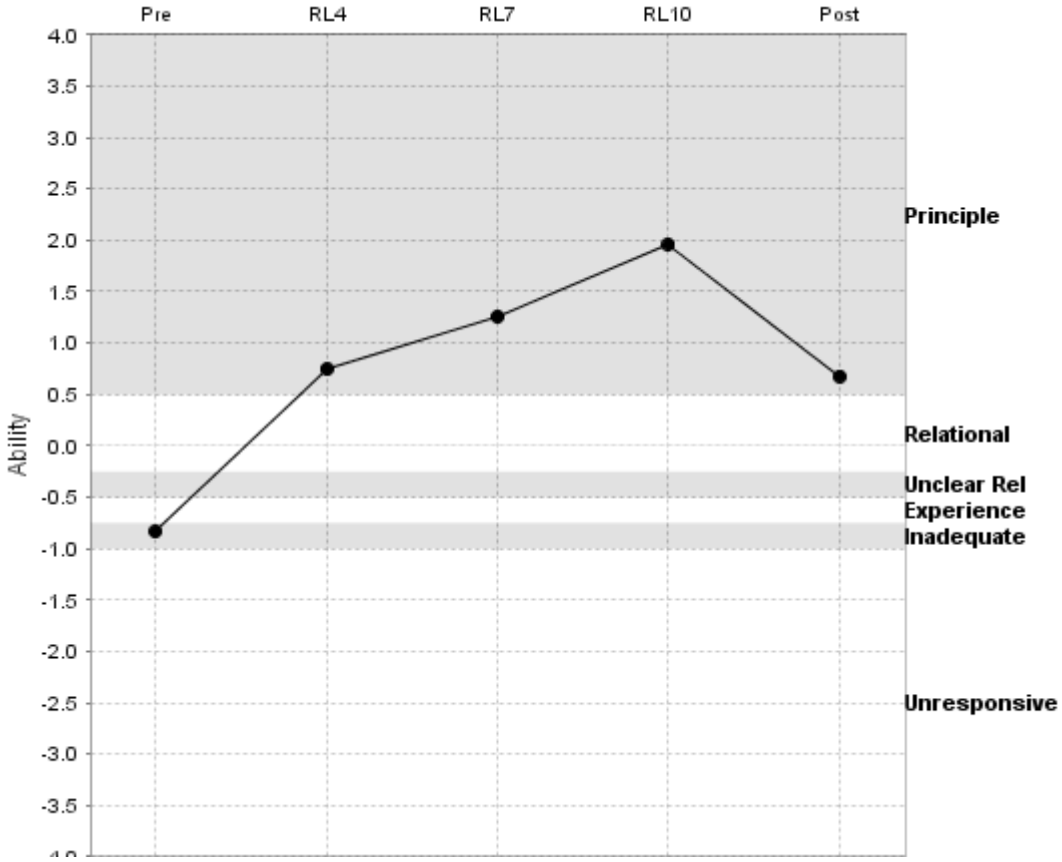

Figure 14. Map of Thurstonian thresholds for Reflective Lesson 4 on the Buoyancy: WTSF progress variable.

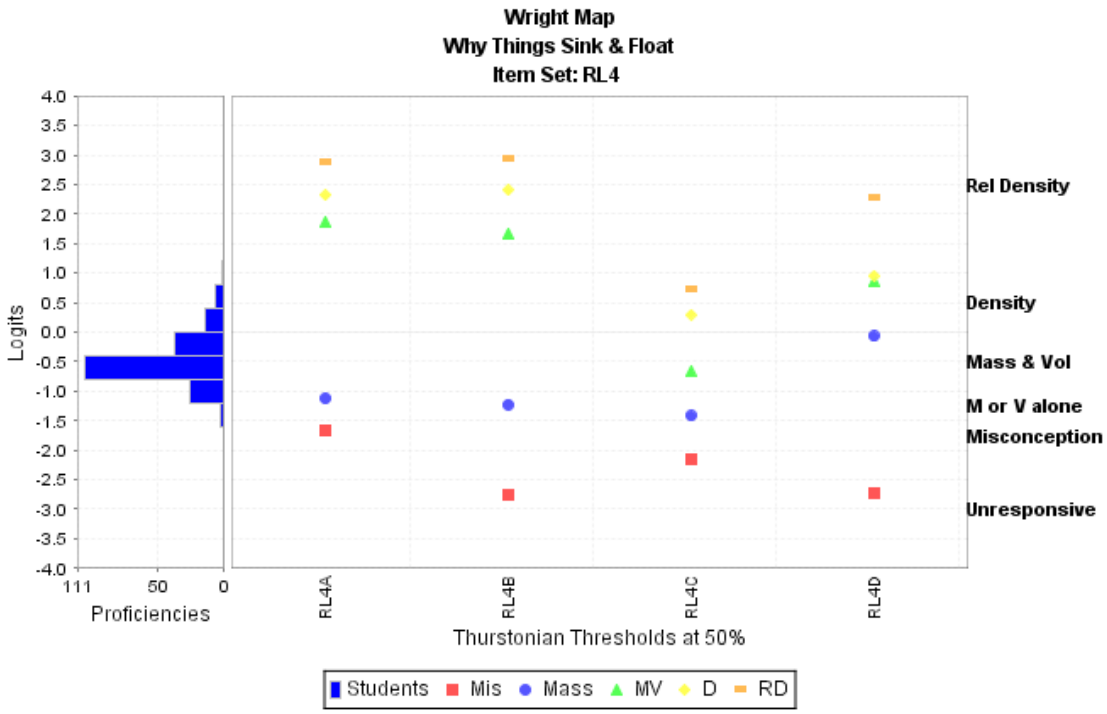

Figure 15. Map of Thurstonian thresholds for Reflective Lesson 4 on the Reasoning progress variable.

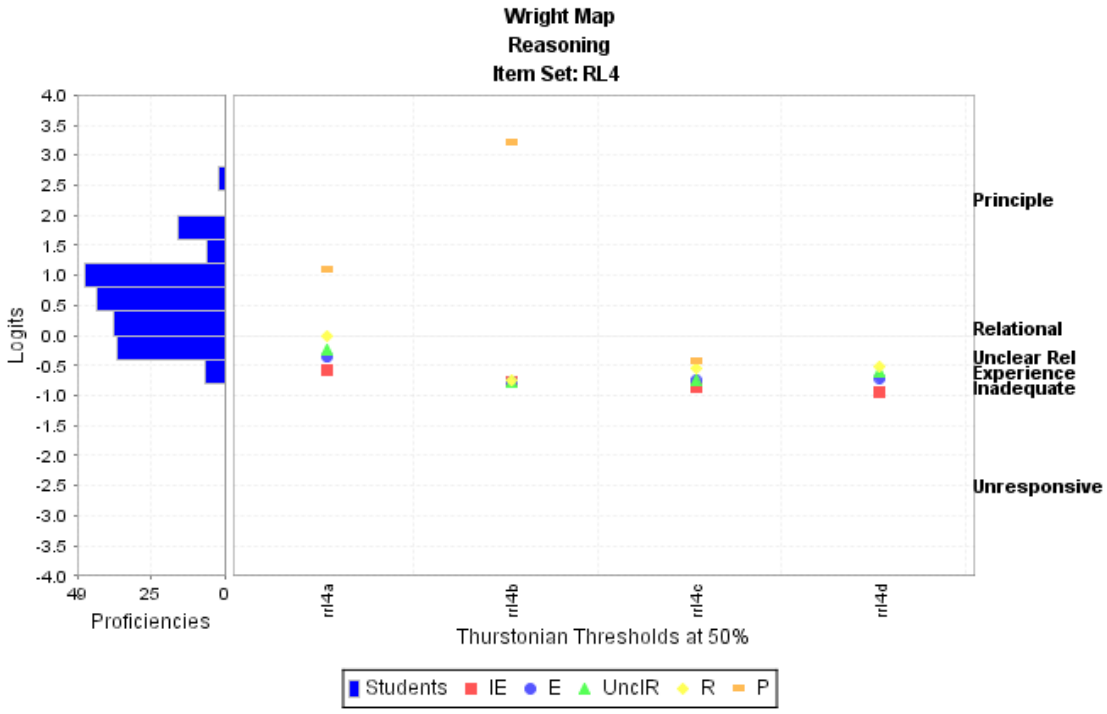

Figure 16. Frequency Map for the WTSF variable for one class after Reflective Lesson 7 is administered. Several students are not operating at the expected level for that point in the unit. The students demonstrating the lowest level of understanding are of the most concern to the teacher.

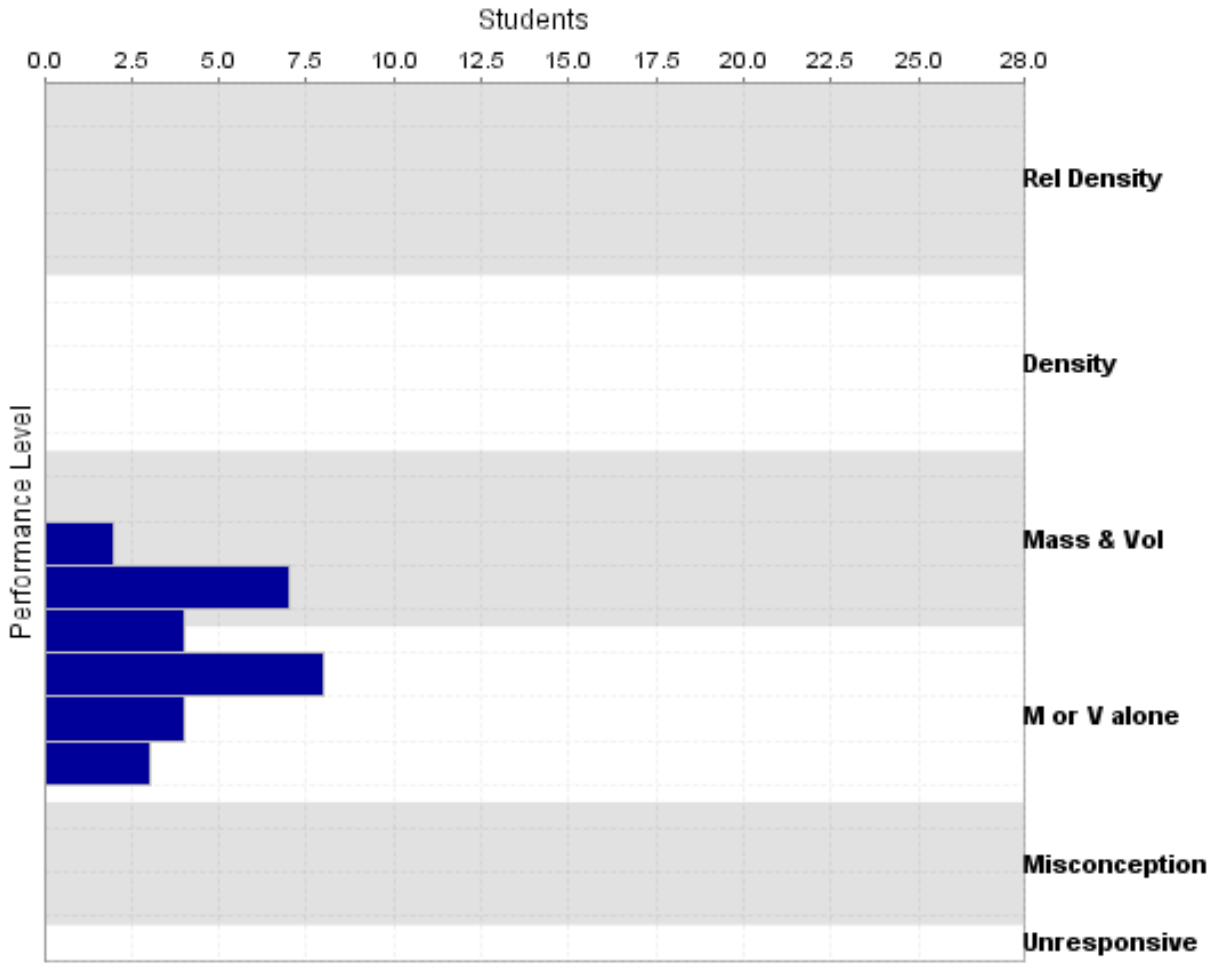

Figure 17. Performance Map on the WTSF progress variable for Amy after completing Reflective Lesson 7.

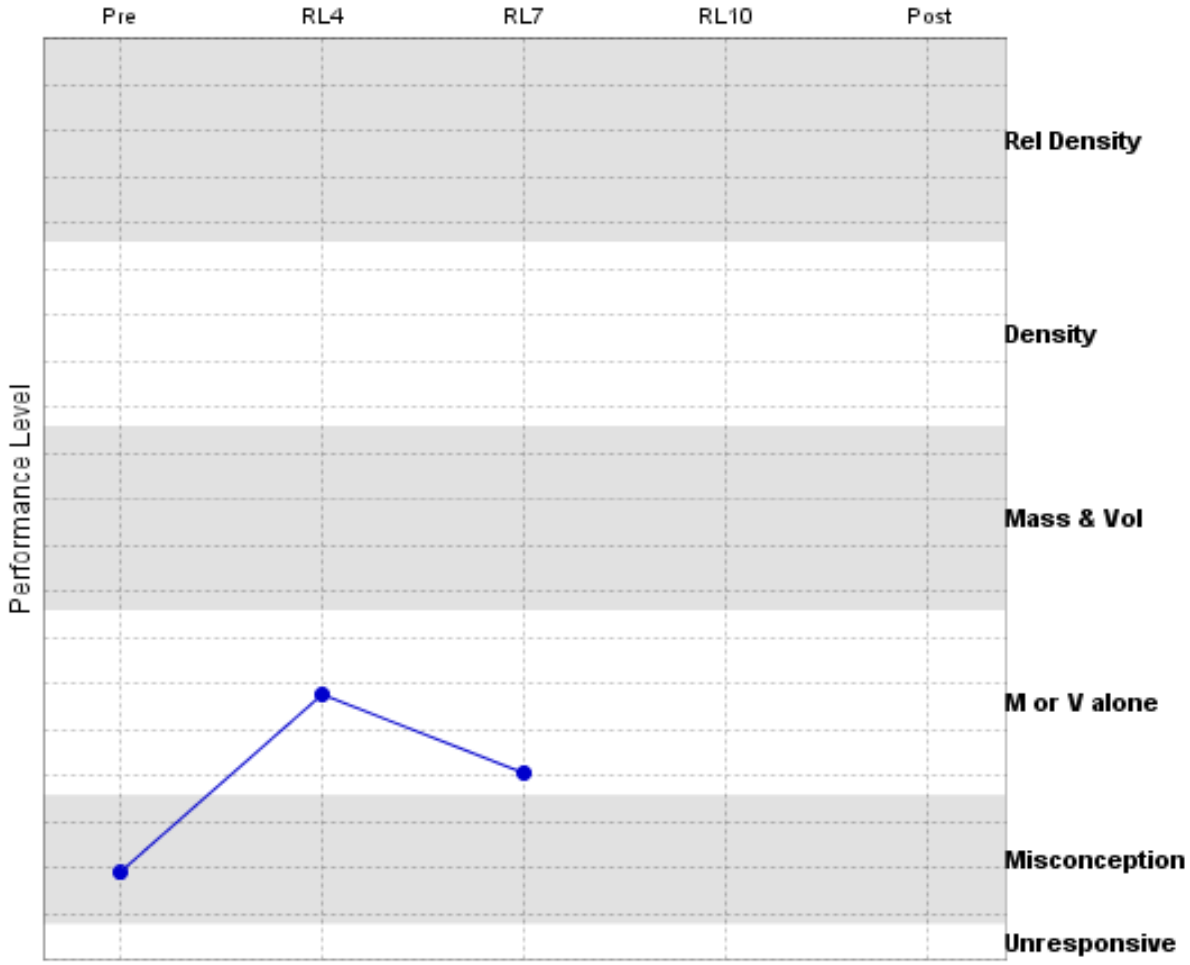

Figure 18. Performance Map on the Reasoning progress variable for Amy after completing Reflective Lesson 7.

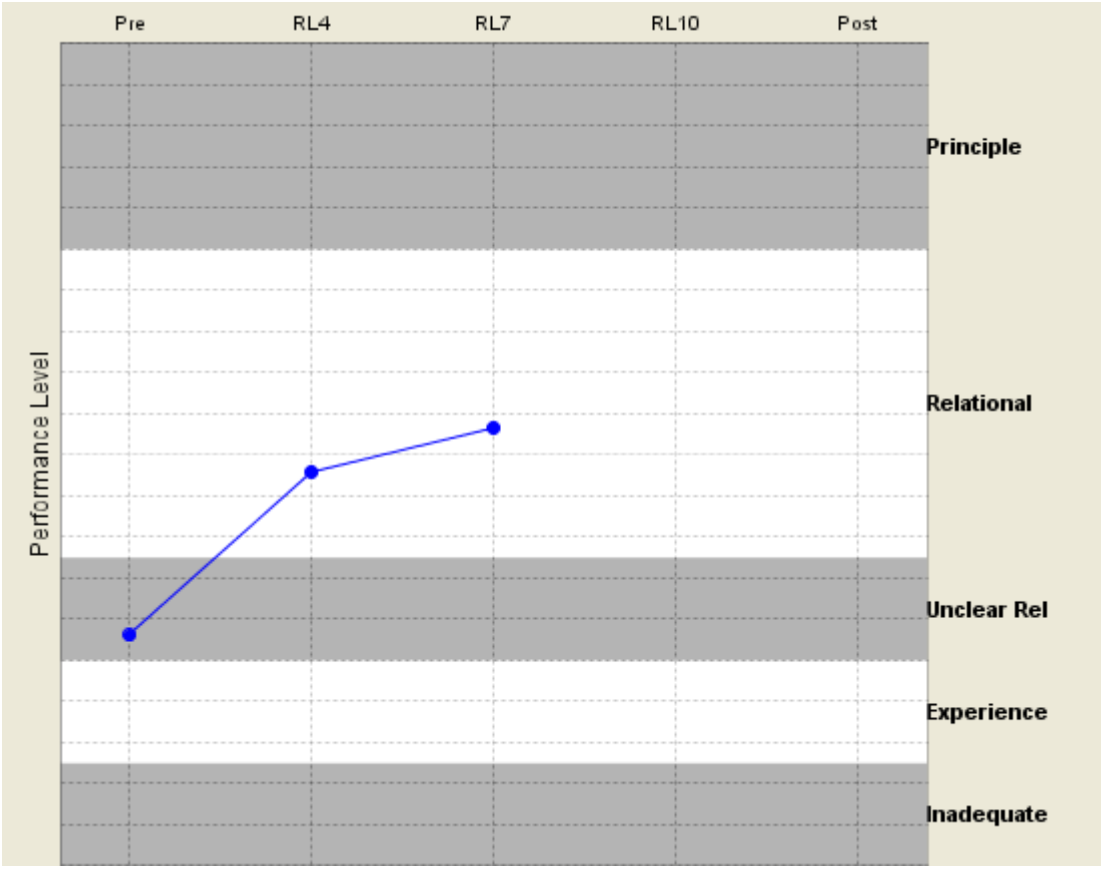

Figure 19. Excerpt of a Wright Map of WTSF for one student on Reflective Lesson 7.