Running Head: PROGRESS VARIABLES

Using Progress Variables to Map Intellectual Development

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

Not to be cited without author's permission.

Publication pending 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, 2006.

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

The current focus on school accountability, accompanied by demands for teacher responsibility for improving student performance, have educators searching for new ways to measure student achievement and growth and 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 intellectual development within a curricular unit was developed to help teachers understand classroom progress and detect individual instructional needs. A process for establishing instructionally useful performance levels that provide an interpretive context for understanding student proficiency and needs is described and teachers' use of graphical representations of student proficiencies for instructional planning is illustrated. An example is provided from Science education based on the topic of buoyancy.

Using Progress Variables to Map Intellectual Development

Introduction

The current focus on school accountability, accompanied by demands for teacher responsibility for improving student performance, have educators searching for new ways to measure student achievement and growth and 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 intellectual development within a curricular unit to help teachers understand classroom progress and detect individual student needs. In this study we found that progress variables can be used to develop an instructionally useful context for understanding student achievement and intellectual growth and that graphical representations of student proficiencies provide useful formative feedback to teachers for planning next steps in the classroom and for identifying specific concepts that individual students need help with.

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. Embedded assessment activities, which are similar to the core hands-on instructional activities of the curriculum, formed the basis for eliciting evidence of the progress variables.

Two progress variables were developed to represent the progression of students toward important curricular goals in the unit; these progress variables reflected student understanding of "Why Things Sink and Float" (the WTSF progress variable) and 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.

Method

An assessment system for the Buoyancy unit was developed 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

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.

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 coverage, which tends to drive the creation of multiple progress variables representing every possible curricular goal, and 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.

Why Things Sink or Float

The first progress variable, initially named Content, was developed to represent the trajectory of learning the science content knowledge in the FAST unit on buoyancy. 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.[1] Figure 1 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 1 about here]

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

The Content progress variable was intended to apply to all items that involved science content knowledge. However, attempts to use this progress variable to analyze student responses from the pilot study revealed that not all of the topics and tasks associated with the curriculum involved the same kind of content knowledge. The progression of understanding represented by the Content progress variable applied most directly to situations dealing with the question of whether an object would sink or float in a given medium, which depends on the mass and volume (and therefore the density) of the object. Many situations in the curriculum, however, deal with the question of how far below the surface a floating object will rest, called the object's *depth of sinking*. Depth of sinking depends on the object's mass and shape (more specifically, cross-sectional area), rather than mass and volume. The Content progress variable proved difficult to apply to students' responses to situations involving depth of sinking, which led to the realization that different trajectories are involved in learning these different concepts. Consequently, the Content progress variable was split into two progress variables, one dealing with the question of Why Things Sink or Float (WTSF) and the other dealing with Depth Of Sinking (DOS). However, while the topic of WTSF is an important theme throughout the curriculum, the topic of DOS is a major focus in only the first several investigations. Consequently, it was decided that the WTSF progress variable would be more useful 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 1. 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 we finally arrived at is shown in Figure 2.

[Figure 2 about here]

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, named Using Evidence, that represented the use of evidence in supporting a claim. Initially, this progress variable was based in large part upon the Using Evidence progress variable developed for the SEPUP curriculum (SEPUP, 1995).

However, analysis of student responses from the pilot study during development of the outcome space (described in the *Outcome Space* section) revealed that students did not use evidence to support their justifications of why things sink or float. Instead, student justifications were based upon things like specific relationships (e.g., “the block is heavy”) or general principles (e.g., “objects with more mass will be more likely to sink”). These observations lead to the development of a new progress variable, called Reasoning, that represents a student’s progress toward increasing sophistication of reasoning displayed in justifications. Designed to satisfy the goals of the project described above, this progress variable (a) deals with a different kind of knowledge than science content knowledge; and (b) is expected to apply broadly to other curricula and assessments, wherever written justifications are required.

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 3.

[Figure 3 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).

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 (Shavelson, SEAL & CRDG, 2005). 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 the addition of written justifications for some items. These additional prompts ask students to explain their choices to the nine items that we felt could be directly related to the WTSF progress variable. This was done primarily to elicit more evidence of student conceptions about why things sink and float. Only the multiple-choice items with justification 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. The multiple-choice items were replaced by the open-ended essay question, "Explain why things sink and float. Write as much information as you need to explain your answer. Use evidence and examples to support your explanation." 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.

Post test

The original post test (i.e. the pre/post test) was replaced by a composite of multiple choice with justification (MCwJ) 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 map 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.

Why Things Sink and Float (WTSF)

A progress guide contains much more detail, and possibly more levels, than the map of the progress variable (Figure 2) 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 4) 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.

Elementary misconceptions were thought to represent an initial understanding of buoyancy based upon everyday experience (for example, with bathtub toys and boats), so responses involving such misconceptions were treated as indicative of a low, pre-instructional level on the WTSF progress guide (labeled "Unconventional Feature").

Incorrect relationships, in which correct concepts are used incorrectly, were thought to represent the first steps in learning to use these new concepts. Consequently, responses involving incorrect relationships were initially treated as sub-levels within the main level on the progress variable dealing with that concept. So, for example, Density was split into three sub-categories, corresponding to: (a) density used incorrectly; (b) density invoked but no relationship specified;

and (c) density used correctly. However, a profusion of sublevels was thought to be potentially overwhelming and a distraction from the main progression represented by the progress variable. As a compromise, a minus sign was introduced so that teachers could indicate a problem with a student's response. For example, a response that used mass incorrectly, claiming that more mass would cause an object to float, could be labeled as M–, indicating the response is at the Mass level but is not fully acceptable. 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.

Imprecise terminology was thought to represent either sloppy use of the correct concept, or use of the correct term before the concept had been learned. These responses were treated as indicative of the level on the progress variable corresponding to the correct (implied) concept, but teachers were again given the option of using a minus sign to indicate that these responses were not fully adequate. As more responses were analyzed, it became apparent that a substantial number of responses invoked things like "size" or "amount" or used "volume" in a way that strongly suggested size or amount (concluding, for example, that a larger volume would be more likely to sink). It was not clear whether these students were thinking about mass, and therefore deserving of an M– label, or whether size was a misconception like shape, and therefore deserving of a UF (Unconventional Feature) label. This tension was resolved by creating a new level specifically dealing with concepts like size and amount, called Productive Misconceptions, located between the Unconventional Feature and Mass levels.

Off-topic responses were thought to represent a very low level of understanding so that the student is unable to respond to the question and instead writes something off-topic.

Consequently, the Off Topic category was positioned as the second-lowest category of the outcome space. Blank responses were assigned the lowest coding category (No Response).

The final WTSF progress guide shown in Figure 4 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 also includes more categories than the number of performance levels defined for the progress variable; we added categories for No Response and Unscorable. 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.

Figure 4 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.

Reasoning

This outcome space was initially developed by a phenomenological process (Marton, 1981; Wilson, 2005) in which student responses to the Reflective Lessons during the pilot test

[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.

were analyzed and grouped into ordered categories. These categories became the qualitative levels of the progress variable and the outcome space. The progress guide for the Reasoning variable is shown in Figure 5. Once again, the progress guide contains more detail than the map of the progress variable (Figure 3), 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, 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 5 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.

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 6. In this example the 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 further 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 6 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 & Setting Cut-Points

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 7 shows on which post test 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.

Figure 7 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

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

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).

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. Once the logit values for the cut-points are established, the interpretive context is set for mapping intellectual development. The standard setting process is described in more detail in Kennedy (2006) and Kennedy and Wilson (2006).

Results

A number of benefits are derived from 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, and estimate a measurement model that maximizes useful interpretation. In this section we address the particular benefits to teachers and their students.

Instructionally useful performance levels are established for each progress variable to provide an interpretive context for understanding student proficiency and instructional needs. 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. 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. 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.

Because performance levels represent curriculum-specific performance based on learning objectives, and these depend on value judgments and expectations, we generally prefer to follow a more subjective approach that relies on the judgments of teachers, curriculum developers, and educational researchers. In this study, we did not have sufficient time to implement that approach. Nevertheless, the quantitative approach used in the current study did produce realistic performance levels that were used to map student intellectual development. An example of a progress map on the WTSF variable for a class in the study is shown in Figure 8. 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 on one instrument 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 8 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 9 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 9 about here

In this study, 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.

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 use primarily graphical means to 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 10), 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 11) shows how a student's proficiency estimate has changed over time. For more detail about student understanding, a teacher might examine a *Wright Map* (Figure 12) for a particular student.

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 10 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 10 about here

Figure 11 shows the *Performance Map* for the WTSF progress variable for one student 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 12, 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 11 about here

Figure 12 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 13 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, 2006; and Kennedy, Wilson & Draney, 2006)

Figure 13 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 indicating particular strengths and weaknesses. Unexpected variations in proficiency or other problems can be noticed early so that corrective steps may be taken while instruction is on-going.

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), 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 a useful approach 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.

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. In X. Liu and W. Boone (Eds.) *Applications of Rasch Measurement to Science Education*. Chicago: JAM Press.

Kennedy, C.A. & Wilson, M. (2006). *Using progress variables to interpret student achievement and progress (BEAR Technical Report 2006-12-01).* Berkeley, CA: Berkeley Evaluation & Assessment Research Center.

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

Marton, F. (1981). Phenomenography: Describing conceptions of the world around us. *Instructional Science, 10,* 177-200.

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.

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).

SEPUP (1995). *Issues, Evidence and You: Teacher's Guide*. Berkeley: Lawrence Hall of Science.

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.

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

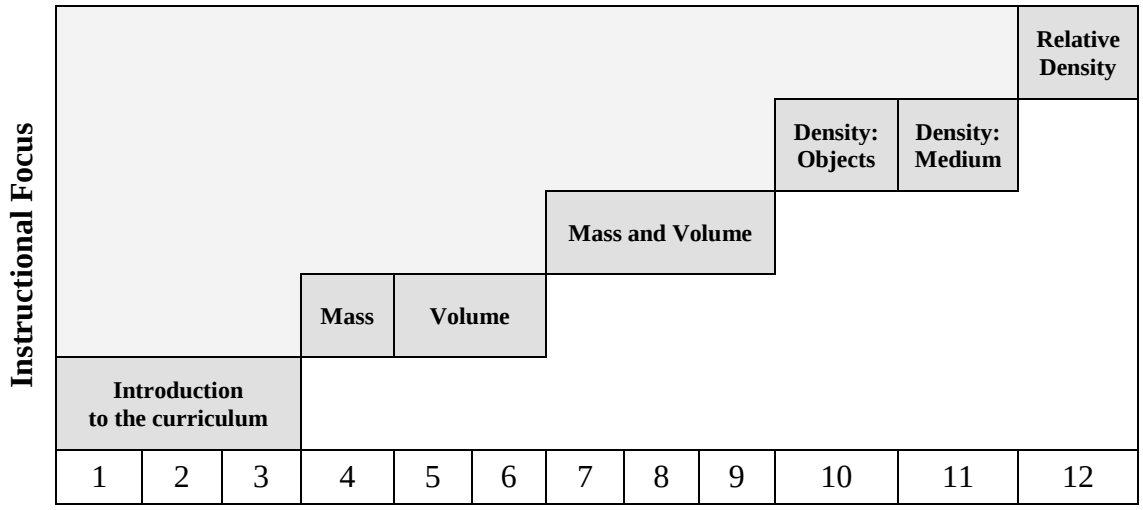

Figure 2. Map of the Buoyancy: WTSF progress variable with qualitative performance levels.

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 3. 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 4. Final version of the progress guide for the Buoyancy: 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 5. 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 6. 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.

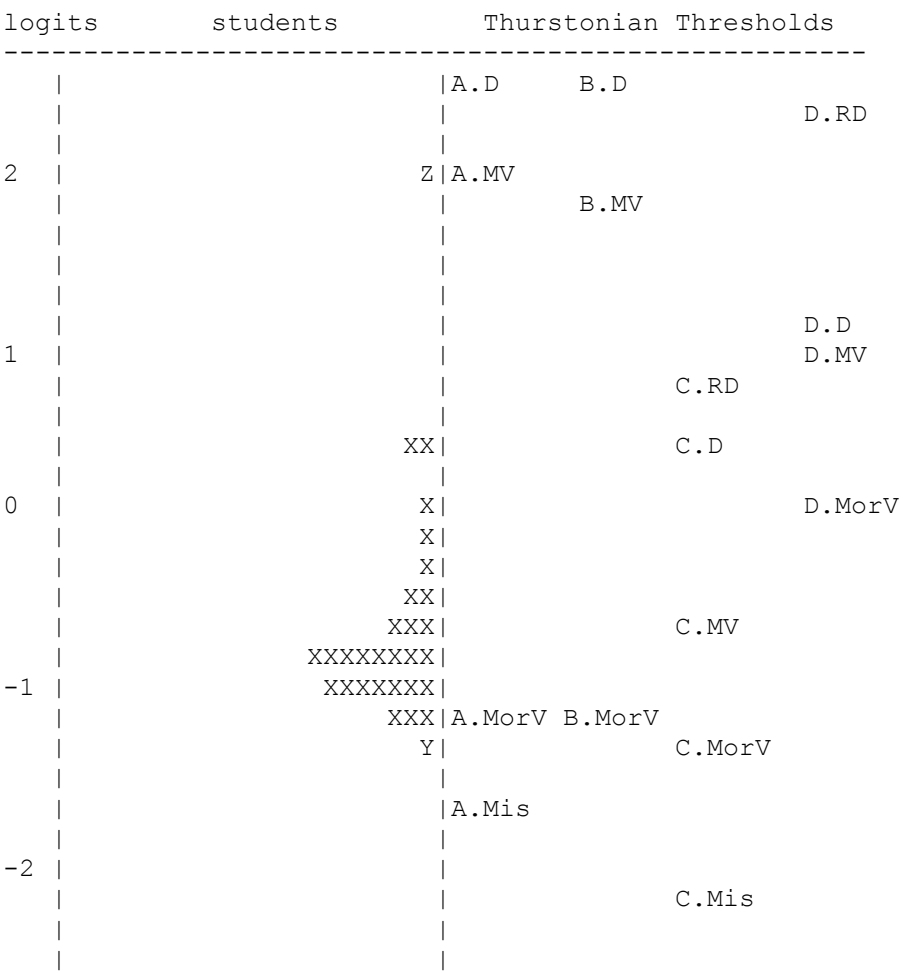

Figure 7. Distribution of pretest and reflective lesson items across five forms of the post test. The numeric value in the Pretest column indicates the number of multiple choice with justification items included on the post test. The shaded areas indicate the form an item was part of. RL4 part C, RL7 part C and RL10 part C were identical items and were calibrated as a single item.

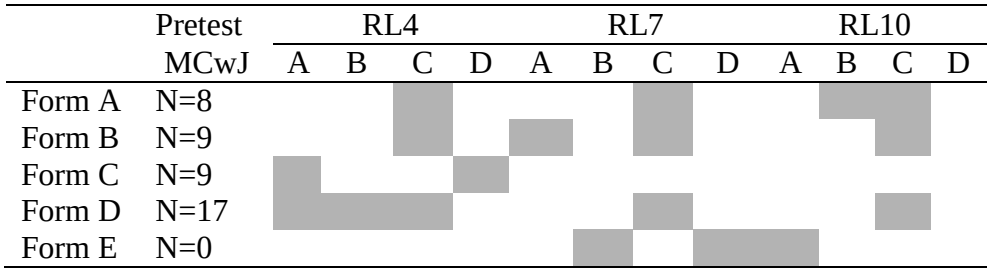

Figure 8. Map of average student progress on the WTSF progress variable for the students of Teacher 3.

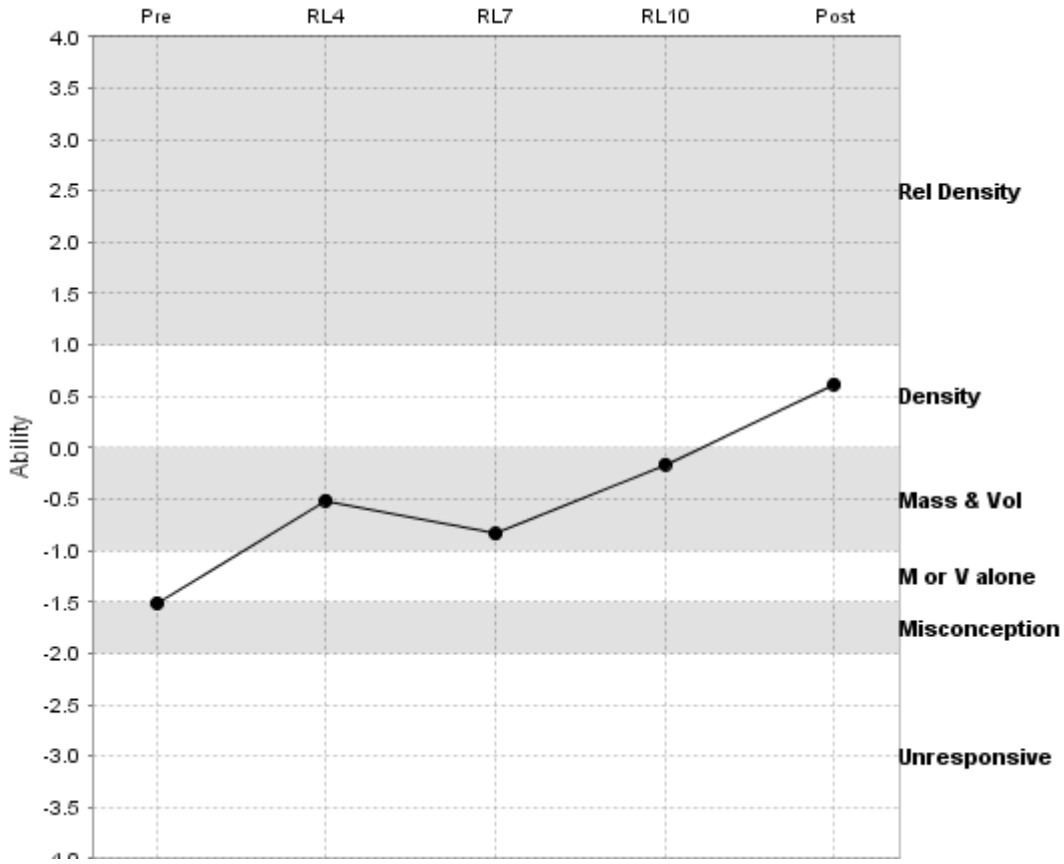

Figure 9. Map of average student progress on the Reasoning progress variable for the students of Teacher 3.

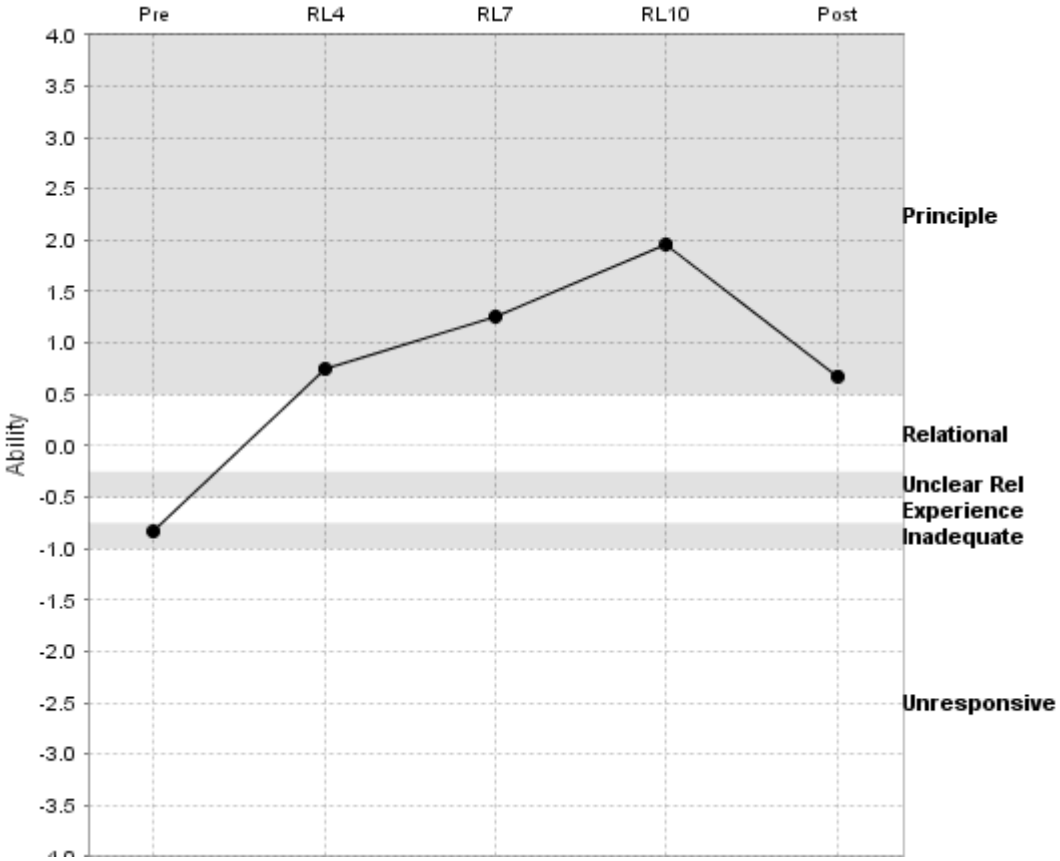

Figure 10. Frequency Map for the WTSF variable for one class after Reflective Lesson 7 is administered. Several students are not operating at the expected level for that point in the unit. The students demonstrating the lowest level of understanding are of the most concern to the teacher.

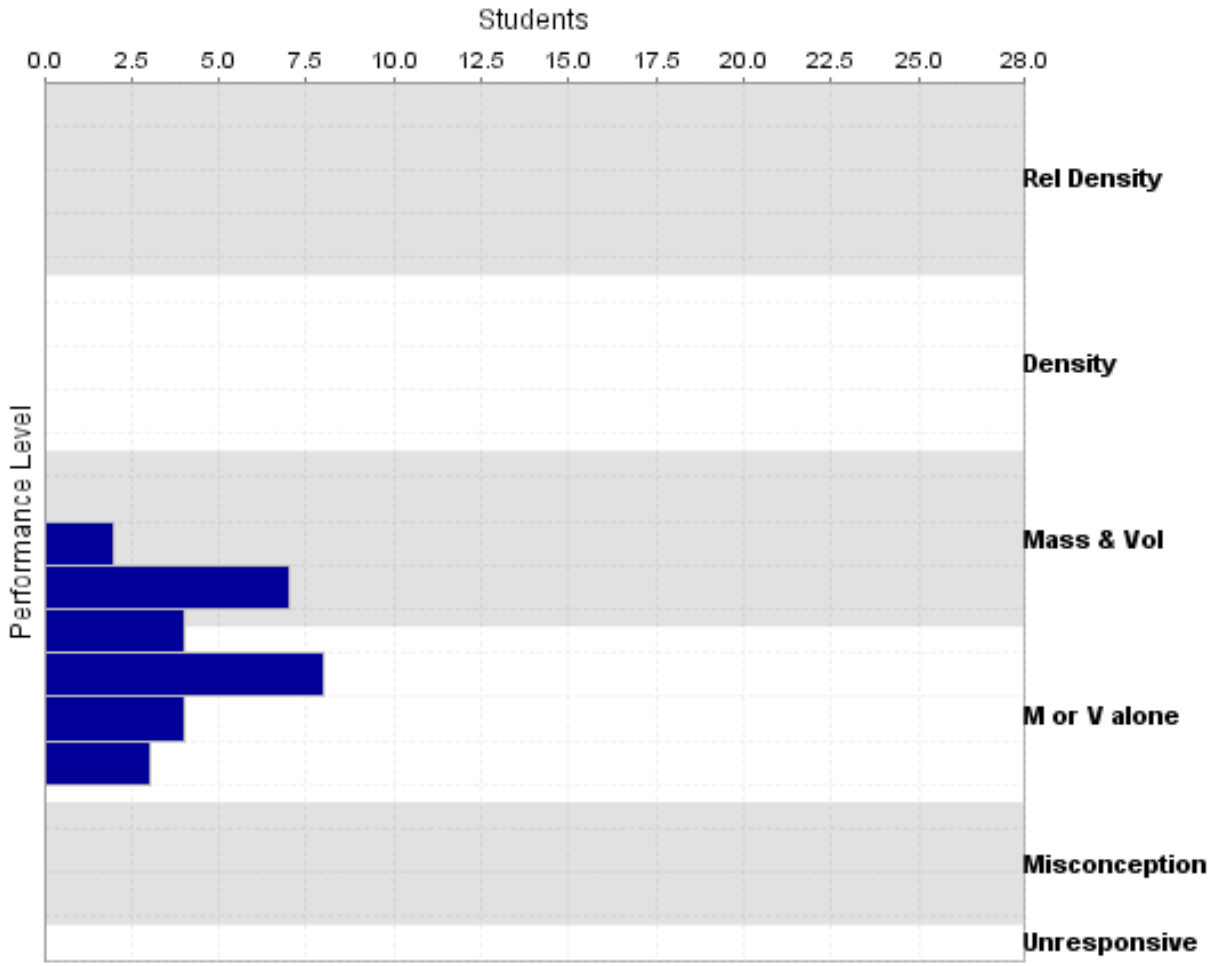

Figure 11. Performance Map on the WTSF progress variable for Amy after completing Reflective Lesson 7.

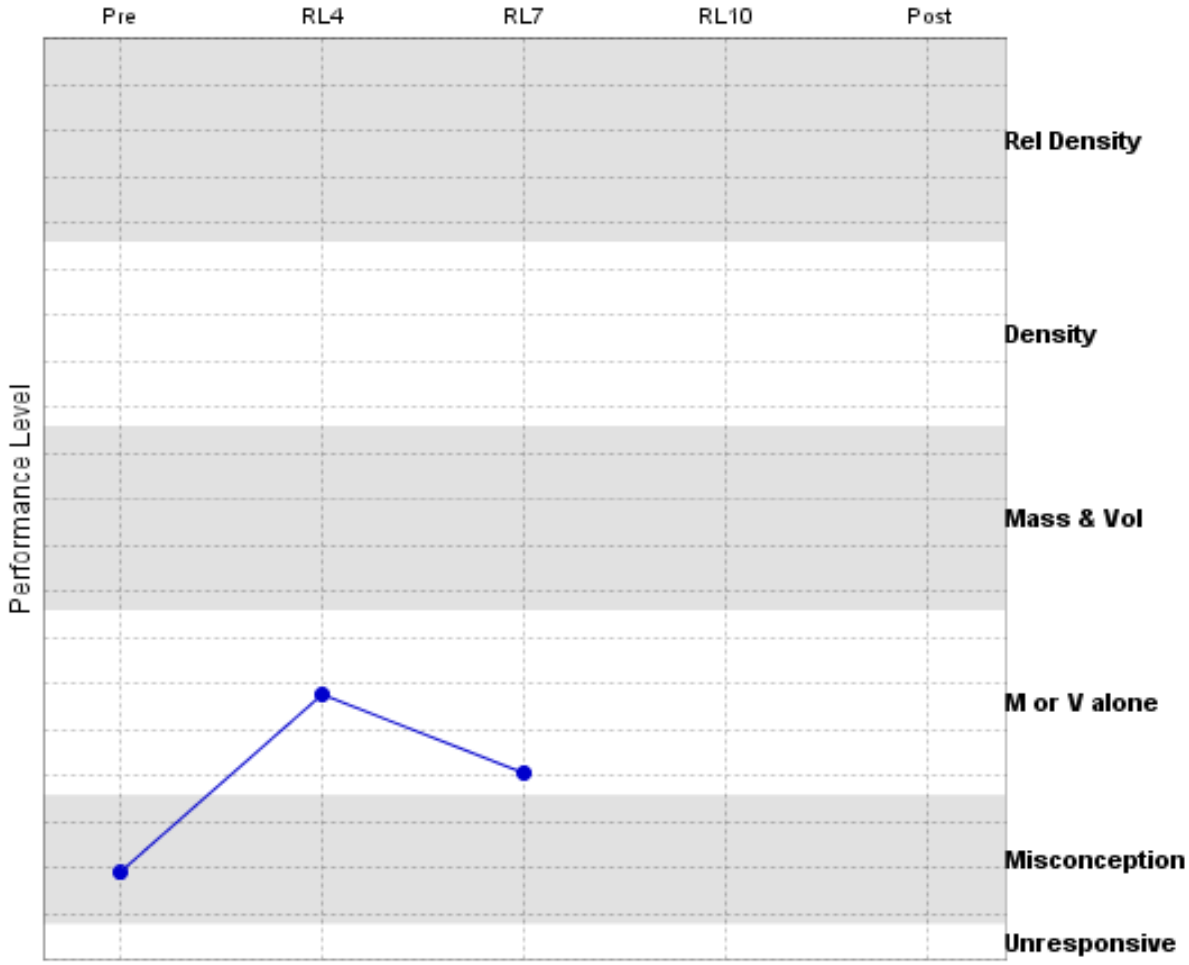

Figure 12. Performance Map on the Reasoning progress variable for Amy after completing Reflective Lesson 7.

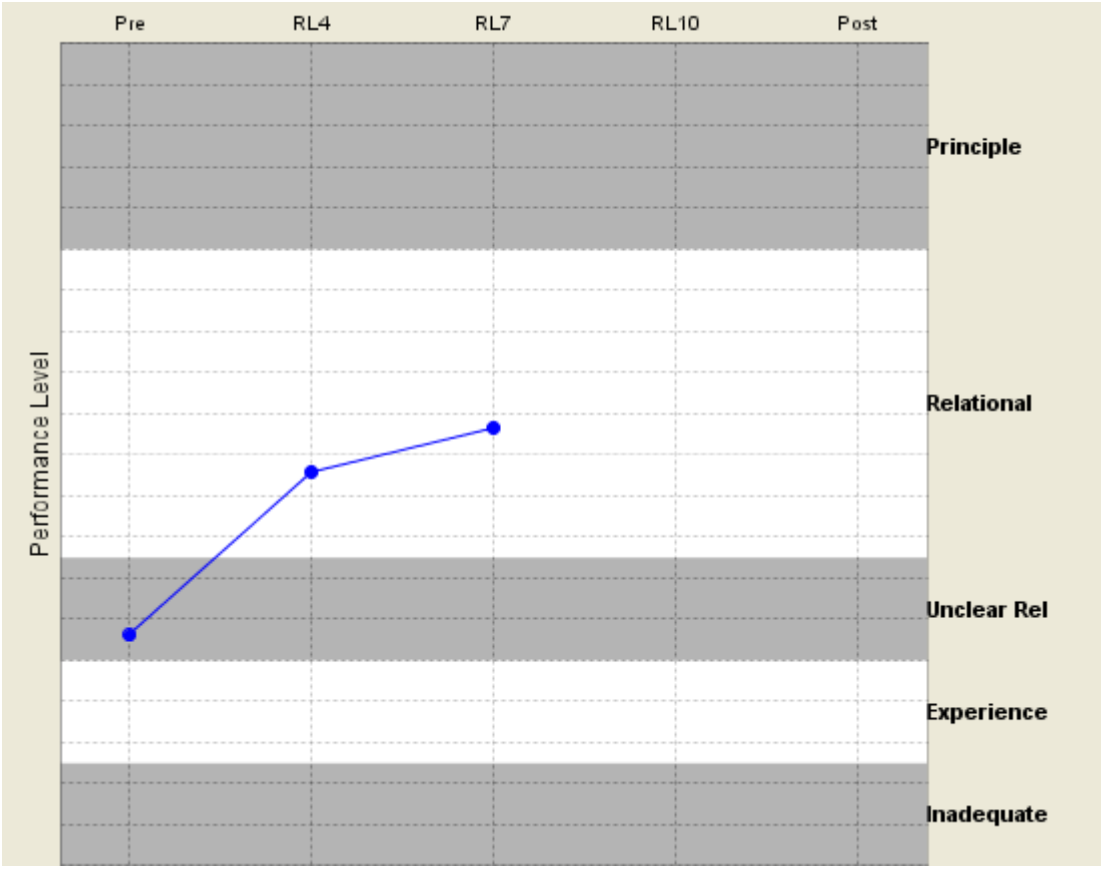

Figure 13. Excerpt of a Wright Map of WTSF for one student on Reflective Lesson 7.