

Simplified Scoring of Performance Activities: Comparing Assessment Stories from Complex and Simple Scoring Approaches

Cathleen A. Kennedy
University of California, Berkeley

Abstract

Advances in assessment theory and measurement techniques have led to new ways of evaluating complex, performance-based activities and drawing useful inferences about student knowledge from such evidence. When these activities require complicated scoring of constructed responses, sometimes on multiple dimensions, the time required to read and evaluate student work may prevent teachers from reaping the benefits. This study compared the information obtained from a generic 3-category scoring approach with that obtained from a 6-category scoring rubric specific to the curriculum. Both models were multidimensional, measuring two aspects of science knowledge from the same set of items, and used multidimensional item response theory (IRT) modeling to fit the polytomous response data. The study employed GradeMap software (Kennedy, Wilson & Draney, 2005) to fit the models and to establish an interpretive context for analyzing proficiency estimates. Psychometric properties of the two models were examined, with both meeting standard IRT assumptions. The findings suggest that scoring student work using the 3-category rubric can reliably reproduce much of the information produced from 6-category scoring. The model was particularly useful for identifying students who were not performing at the targeted level and in providing information regarding the concepts those students would benefit from focusing on right away.

Keywords: IRT, Rasch, IRM, MIRM, MRCML, multidimensional, measurement, scoring, assessment, GradeMap

The author gratefully acknowledges the contributions of Nathaniel J. S. Brown, Karen Draney, Diana J. Bernbaum, Lydia Ou Liu and Xiaohui Zheng. 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.

Simplified Scoring of Performance Activities: Comparing Assessment Stories from Complex and Simple Scoring Approaches

Introduction

The benefits of meaningful scoring rubrics have been widely lauded by the educational and research communities (Adams, Wilson, & Wu, 1997; National Research Council, 2001a, 2001b), but implementing such approaches in the classroom can be daunting for teachers (Herman, Osmundson, Ayala, Schneider, & Timms, 2005; Howe, 1997). Teachers are being asked to embed formative assessment activities into their regular classroom practice, perhaps as often as every week, to better track student progress and diagnose student needs. Ideally, feedback to students should happen as soon as possible, while students are still learning the concepts that were assessed (Black & Wiliam, 1998a, 1998b). When these activities require complicated scoring of constructed responses, sometimes on multiple dimensions, the time required to read and evaluate student work may prevent teachers from reaping the benefits that could be obtained from the additional information.

Advances in assessment theory and measurement techniques have made it possible to model complex multidimensional trajectories of student change. We can use these new techniques to design rich assessment tasks that elicit evidence of such change, and to construct measurement models to draw valid and reliable inferences from that evidence. Between the delivery of these assessment tasks to students and the calculation of proficiency estimates lies the process of transforming student work into the evidence from which these inferences can be drawn. Quite often, particularly with performance-based activities, that evaluation process cannot be automated, nor do we want it to be. Reading and evaluating student work is not only useful for the formative feedback it can generate for students; teachers also benefit when they reflect on how they might improve the alignment of instruction with curricular goals and on techniques they might use to develop their students' metacognitive skills (Herman, Osmundson, Ayala, Schneider, & Timms, 2005).

This study proposes that by starting with a detailed scoring rubric describing responses that differentiate several levels of knowledge, one can align individual items to specific levels on that rubric, and then use a simpler three-point scale to score student work. The three scores simply indicate whether a response meets the targeted level, misses it, or exceeds it. If this simplified scoring approach provides nearly as much useful formative information as the more complex approach, teachers will have a technique to expedite the evaluation of student work and provide meaningful feedback to students in a more timely manner.

We are mindful here of distinguishing between *psychometrically optimal* assessment data and *practically useful* data. Assessment data are the outcome of a complete assessment delivery system as defined by Almond and his colleagues as a four-process model (Almond, Steinberg, & Mislevy, 2002) that includes administering the assessment, gathering responses, scoring the responses, and summarizing and analyzing the response scores. The concept of data-driven decision-making is tied to the capability of teachers (and students) to use assessment data to draw inferences about what each student knows relative to curricular goals and to take action based upon those inferences (NSSE, 1998). Practically useful data provides reliable inferences

about student knowledge that can inform the decisions teachers and students make about instructional “next steps.”

The three-level scoring rubric considered in this study could be applied to assessment items in many domains. Here, it is used to evaluate science content knowledge and science inquiry skills. The rubric is applied to both embedded assessment activities that are, to students, a normal part of day-to-day classroom activities, and to a summative end-of-unit test. Use of the rubric requires an underlying concept of learning development that can be described in distinct categories of responses that progress from exhibiting “less” to “more” knowledge. In addition, the levels must be linked to specific instructional units. This is necessary so that individual items can be aligned with targeted learning objectives along that continuum and student work can then be interpreted to provide feedback that has a specific and direct impact on learning (Kennedy, 2005). The interpretation of students’ scores is facilitated by the use of a software program, GradeMap (Kennedy, Wilson, & Draney, 2005), that produces graphical reports of student status and progress using multidimensional item response method (MIRM) proficiency estimates.

This paper compares the assessment stories that can be told using a complex multi-level scoring approach in which responses to each item are assigned a score along the entire rubric with an approach that evaluates responses as simply “on target,” “needs help,” or “advanced.” We think of an assessment story as a description of what a particular student (or a group of students) knows and can do at a particular point in time and what this implies about next steps for the student (or students). The quality and usefulness of that story depend upon the kinds of inferences that can be drawn from the evidence. The learning theory, the measurement model that operationalizes that theory, and the validity and reliability of the resulting measures all contribute to the quality of the assessment story. Each of these factors is explored for the two scoring approaches in this study.

Data

This study is part of an NSF-funded Centers for Assessment and Evaluation of Student Learning (CAESL) project (NSF grant ESI-0119790; Kennedy, Brown, Draney, & Wilson, 2005) including researchers from the University of California at Berkeley, UCLA, Stanford University, and WestEd. Constructs, formative and summative assessment activities, and scoring guides, or rubrics, were developed by project team members for use in a unit on Buoyancy from the *Foundational Approach to Science Teaching (FAST) Physical Science* curriculum (Pottenger & Young, 1992). We refer to the project as “CAESL/FAST” in this paper.

Five assessments were administered in the CAESL/FAST project: a pretest, three embedded formative assessment activities, and a post test. The embedded activities were referred to as “Reflective Lessons” in the curriculum and are identified as “RL4,” “RL7,” and “RL10” in this study. The pretest data were not included in the current study because a pretest is not usually administered with the curriculum; our intention in the current study is to evaluate the use of different scoring rubrics in a standard classroom implementation of the curriculum. The embedded assessment activities were in-class lab activities; the assessment component involved students answering open ended questions about their experience as they worked through the labs. Each included a graph interpretation item, a predict-observe-explain item, an essay explaining

why objects sink or float, and a challenge predict-observe item that addressed concepts to be taught in the next lesson.

Data were gathered from students in thirteen California middle schools in the 2003-04 school year. Eight of the original thirteen teachers were able to provide complete data from all of the summative and formative assessments, for a total of 194 students for this study.

Methods

Scoring

Four project researchers scored all of the student work for this project. They met periodically to establish standards for scoring responses in a consistent manner. For the current study, we assume that raters scored consistently. A study of inter-rater reliability is currently under way.

Two scoring guides were used for this study, the first a six-category rubric for scoring responses dealing with content knowledge about buoyancy, referred to as the “Why Things Sink and Float” (WTSF) progress variable, and the second a six-category rubric for scoring the type of reasoning students exhibited in their responses, referred to as the “Reasoning” progress variable. These two scoring guides are shown in Figures 1 and 2. Each row contains information about a level of knowledge or skill, from less sophisticated at the bottom to more sophisticated at the top. The left column contains the scores teachers assign to responses, the center column contains

Progress Variable: Why Things Sink and Float				
Level		What the Student Knows		Example Responses
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.”
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.”
M	V	Mass Student knows that floating depends on having a small mass.	Volume Student knows that floating depends on having a large volume.	“An object floats when its mass is small.” “An object floats when its volume is large.”
MIS		Misconception Student thinks that floating depends on having a small size, being flat, filled with air, or having holes.		“An object floats when it is small.” “An object floats when it is flat.”
OT		Off Target Student does not attend to any property or feature to explain floating.		“I have no idea.”

Figure 1. Six-category scoring guide for the *Why Things Sink and Float* progress variable for the CAESL/FAST curriculum.

Progress Variable: Reasoning to formulate an explanation		
Level	What the Student Knows	Example Responses
P	Principled Student uses an explicit principle that applies to objects in general.	“An object floats when its mass is small.”
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.	“It floats because its mass is small.”
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.	“It floats because of its mass.” “It floats because it has less.”
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 I have a bar of soap like that at home and it floats.”
IE	Inadequate Explanation Student either restates their answer as an explanation, or simply asserts that their answer is correct.	“It floats because it floats.” “It floats because I know it does.”
OT	Off Target Student cannot or does not give an explanation for their answer.	“I have no idea.”

Figure 2. *Original scoring rubric for the Reasoning progress variable for the CAESL/FAST curriculum.*

descriptions of the knowledge exhibited at that level, and the right column provides an example of a response at that level. The same guides were used to evaluate all items.

The researchers evaluated each response and assigned a score from each of the two scoring guides, one score on the WTSF progress variable, and one on the Reasoning progress variable. Each score was treated as an independent variable; a future study will evaluate the extent to which the WTSF and Reasoning scores on each item were conditionally dependent. Blank or illegible responses were treated as missing data in this study because they contribute no information to our understanding of student proficiency. In classical testing practice, these responses are typically treated as incorrect responses.

To generate the score data for the three-category model we aligned each WTSF item with a targeted level on the WTSF six-category scoring rubric, shown in the right-hand column of Table 1. The targeted level was selected by evaluating the content of the items. The alignment shown in Table 1 indicates that items “20,” “21,” “23,” and “25” target responses that would indicate student understanding of relative density. Note that none of the items directly targets responses that indicate underlying misconceptions or that are off target.

Table 1.
Alignment of items to six-category score levels for the Why Things Sink and Float progress variable of the CAESL/FAST curriculum.

6-category rubric	Items targeting the 6-category level
Rel. Density	20, 21, 23, 25
Density	RL4C, RL7C, RL10A, RL10B, RL10C, RL10D, 14, 16, Part4
Mass & Vol.	RL7A, RL7B, RL7D, 6, 10
Mass	RL4A, RL4B, RL4D, 13, Part2, Part3
Mis.	N/A
Off Target	N/A

We then converted the six-category scores into three-category scores using “x” to represent responses at the targeted level, “-“ to represent responses scored below the targeted level, and “+” to represent those above the targeted level. Transforming scores in this manner is straightforward in the GradeMap software; recoding patterns can be defined on the fly for individual items. Since none of the items targeted the original “Off Target” level, every item had a potential “-” score; however, some items targeted the “Relative Density” level, so some items did not have a potential “+” score. For the Reasoning variable, we designated a “Relational” level response as the targeted level for all items. Then, the “Principled” level was scored as “+” and all the other categories were scored as “-.”

Calibrating the Assessments

The psychometric approach we used to model the response data is a multidimensional item response model known as the multidimensional random coefficients multinomial logit model (MRCML; Adams, Wilson, & Wang, 1997). Both test and item multidimensionality are accommodated in the model, and both dichotomous and polytomous scores can be used.

We applied Master’s Partial Credit Model (1982) to our data. This Rasch-family model provides a convenient way to develop estimates of person proficiency and item difficulty using the same scale. When the model provides a good fit to the data, the interpretation of person proficiency estimates relative to item difficulties at each response level is a powerful formative assessment approach. We began by calibrating the items from the post test and examining the psychometric properties of that test. To represent the theoretical model of what the assessments were intended to measure, a two-dimension partial credit model was fit to the data.

To make valid comparisons of student proficiency at different points in time from different assessment tasks, it is necessary to establish the relative difficulties of items across all of the assessments. The post test was designed to contain two or more items in common with each of the other assessments. To scale these together, we anchored the common items on the other assessments with the item difficulties obtained from the post test calibration. Then, each assessment was calibrated individually.

After calibrating, we examined each model first for fit, essentially ensuring that each met the standard assumptions of IRT models, second for the quality of the information generated for persons, and finally for interpretive consistency. If either of the models violated basic assumptions of IRT measurement, then the proficiency estimates generated from that model

could not be deemed reliable enough to use for formative feedback. If the quality of information differed substantially between the two models, then one model might be preferred on that basis alone. And, given that the assumptions were met, if the three-category model provided essentially the same instructional advice for individual students as the six-category model, then that model would be preferred as the most parsimonious and easiest to implement.

Criterion Zones

The first step we took to establish a foundation for interpreting proficiency estimates was to define zones along the progress variables to identify qualitatively distinct levels of knowledge that students were expected to encounter as they progressed from less sophisticated to more sophisticated understanding. The specific ranges for each score level along a progress variable are called “Criterion Zones” in GradeMap. The values can be computed from the Thurstonian thresholds obtained during calibration of the instruments. 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. A threshold of 1.25 for step 2 of an item means that a person with a proficiency of 1.25 is equally likely to provide a response that will be scored in category 2 or above or a response below category 2.

A criterion zone is defined for each response category of a progress variable. As shown in Figure 3, the cut-point between two zones is the midpoint between the means of the Thurstonian thresholds of the two zones. Note, however, that Thurstonian thresholds are not computed for the lowest category, so there is no mean 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.

Constructing Criterion Zones for the 6-Category WTSF Progress Variable

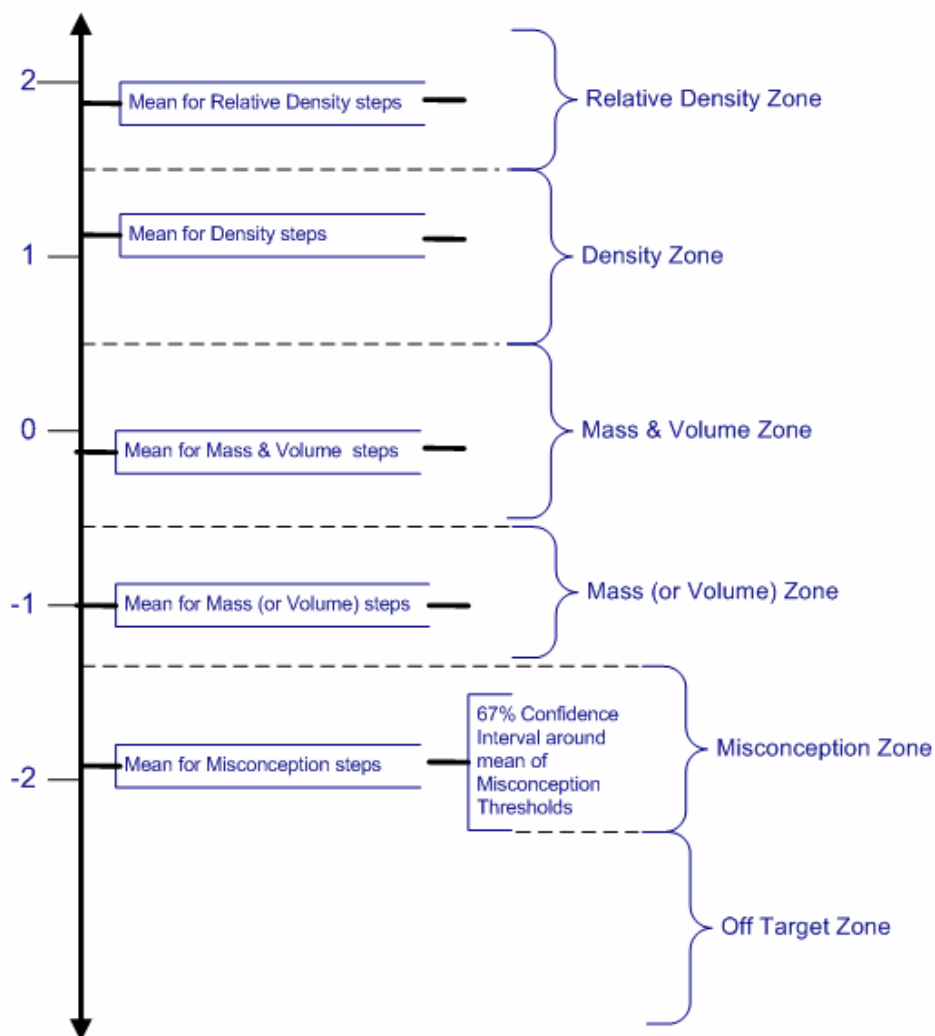


Figure 3. Means of the Thurstonian thresholds and cut-points for the criterion zones for the WTSF progress variable for the six-category model from the CAESL/FAST curriculum.

A second aspect of establishing a context for interpreting proficiency estimates is defining the learning progression anticipated by the curriculum. A level of performance is selected for each assessment and the midpoint of the criterion zone for the targeted level is applied as the expected proficiency for that assessment. For example, when completing assessment activity “RL4,” students are expected to be using the concept of mass in explaining why objects sink or float. The midpoint of the criterion zone for the “Mass” category under the six-category model would be used as the targeted proficiency level for RL4. The highest criterion zone does not have an upper bound, so we use the mean of the Thurstonian thresholds of that response category.

These zones and expected proficiencies are criterion-referenced expectations, based upon the content of assessment items relative to the curricular goals as they have been targeted by the instructional program.

Proficiency Estimates

The GradeMap program allows the user to select expected a-posteriori (EAP), maximum likelihood, or plausible value estimates of person proficiency. We used EAP estimates for this study. These are computed from conditional response probabilities to which a multivariate prior distribution is applied. The EAP estimate is the mean of the posterior distribution of proficiency for the person, with an estimate produced for each progress variable. We did not use maximum likelihood estimates because they do not take the covariance of the two progress variables into account.

A proficiency estimate is interpreted in conjunction with item difficulties. In the case of polytomous items, we examine Thurstonian thresholds for each response level. GradeMap produces *Wright Maps* to graphically align proficiency estimates with item difficulties expressed as Thurstonian thresholds. An example is shown in Figure 4. Person proficiencies are displayed on the left side, while item difficulties are displayed on the right. The entry “13.MV” is interpreted as a response scored in the MV category on item 13.

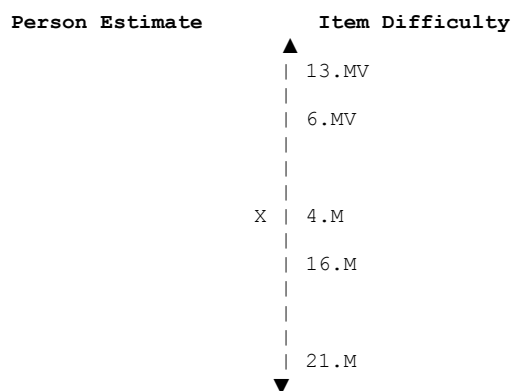


Figure 4. *Hypothetical Wright Map for one respondent and five items.*

This hypothetical example indicates that the probability of a person with a proficiency estimate at the X location responding at the “M” (mass) level or higher on item 4 is 0.5. The probability of that person responding at the “M” level or higher to items 16 and 21 is greater than 0.5, and the probability of responding at the “MV” level to items 6 and 13 is less than 0.5.

The Six-Category Model

The underlying theory of the six-category model is that students in the curriculum progress from bringing nuanced understandings about why things sink or float to understanding that two properties of matter, mass and volume, play a role. They then develop an understanding that it is the *relationship* of mass to volume that affects floating, and then to an understanding of the concept of density as a property of matter that affects floating. The final stage is an understanding that it is the relationship of the density of an object to the density of the medium it is in that determines whether a particular object will sink or float (Kennedy, Brown, Draney, & Wilson, 2005).

These qualitatively distinct levels of understanding are described in a *Construct Map* configured in the GradeMap software. Initially, the proficiency range is simply subdivided into equal zones

for each response level. More precise criterion zones were identified after the data were calibrated, following the procedure described in the *Criterion Zones* section. Figures 5 and 6 show the final GradeMap *Construct Maps* for the six-category WTSF and Reasoning progress variables, which parallel the information contained in the scoring guides displayed in Figures 1 and 2, with the addition of a logit scale to indicate the magnitude of each criterion zone.

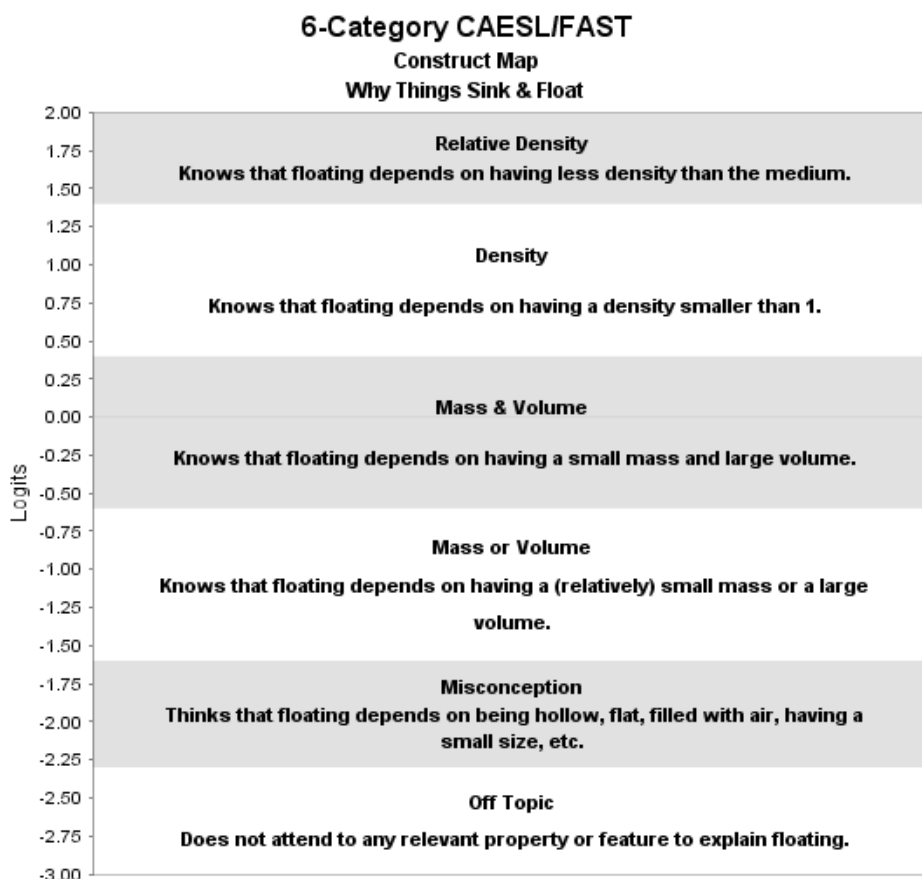


Figure 5. *Construct Map from GradeMap software for the six-category Why Things Sink and Float progress variable of the CAESL/FAST curriculum.*

Criterion zones for the Reasoning variable were somewhat problematic: the means for the Experience, Unclear Relational, and Relational categories were very close together. Note in Figure 6 that the resultant criterion zone for the “Unclear Relational” level is extremely narrow. Once established, these criterion zones are used in numerous GradeMap reports of student status and progress.

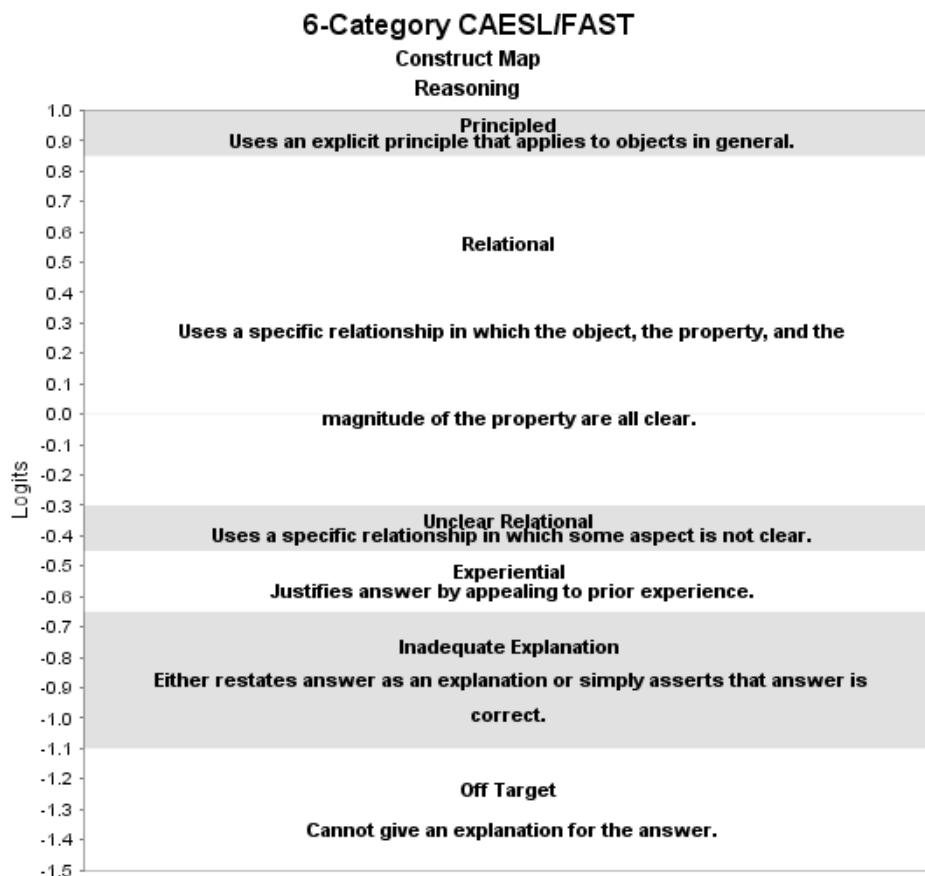


Figure 6. Construct Map for the six-category Reasoning progress variable of the CAESL/FAST curriculum.

The expected proficiency levels for each instrument are defined by the midpoint of the criterion zone of the level targeted by the instrument on each progress variable. Table 2 shows the alignment of assessment instruments with targeted levels on the WTSF and Reasoning progress variables. On the WTSF variable, RL4 targets an understanding of how mass affects floating and sinking, so the expected proficiency for that instrument is the midpoint of the “Mass” criterion zone. For RL7 we use the midpoint of the “Mass & Volume” zone, and for RL10 we use the midpoint of the “Density” zone. Since a midpoint cannot be identified for the “Relative Density” zone, we use the mean of the Thurstonian thresholds for that category as the expected proficiency of the post test.

Table 2.

Alignment of assessments to targeted six-category response levels on the WTSF and Reasoning progress variable of the CAESL/FAST curriculum.

Instrument	WTSF level	Reasoning level	
RL4	Mass alone	-1.0	Relationships -0.2
RL7	Mass & Volume	-0.1	Relationships 0.5
RL10	Density	1.1	General Principles 1.9
Post test	Relative Density	1.9	General Principles 2.3

The Reasoning progress variable was never directly targeted by the curriculum, however students were periodically asked to predict whether a particular object would sink or float, and then to explain why they made that prediction (these activities were called POEs for predict-observe-explain). For this study, we hypothesized how teachers targeted instruction about constructing explanations based upon how student responses actually changed over time. We found that most students used relationships in their explanations in all of the assessments, but some students started transitioning to using general principles at RL10. As shown in Table 2, we theorized that RL4 and RL7 targeted the Relational level, while RL10 and the post test targeted the Principled level. When two assessments targeted the same response category, we used the mean of the Thurstonian thresholds for the first assessment's target level, and the upper limit of the 67% confidence interval of the thresholds for the second level as a convenient way to indicate further progress. The values are displayed in Table 2 under the Reasoning Level columns. Figures 7 and 8 illustrate the resulting expected learning progressions on the WTSF and Reasoning variables for the six-category model.

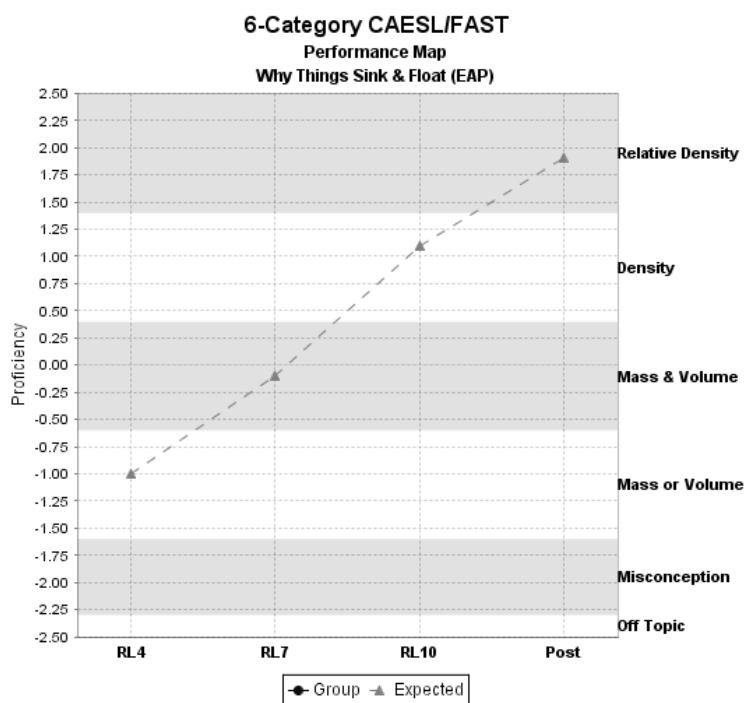


Figure 7. Performance Map showing progression expected by the curriculum for the WTSF variable of the CAESL/FAST curriculum for the six-category model.

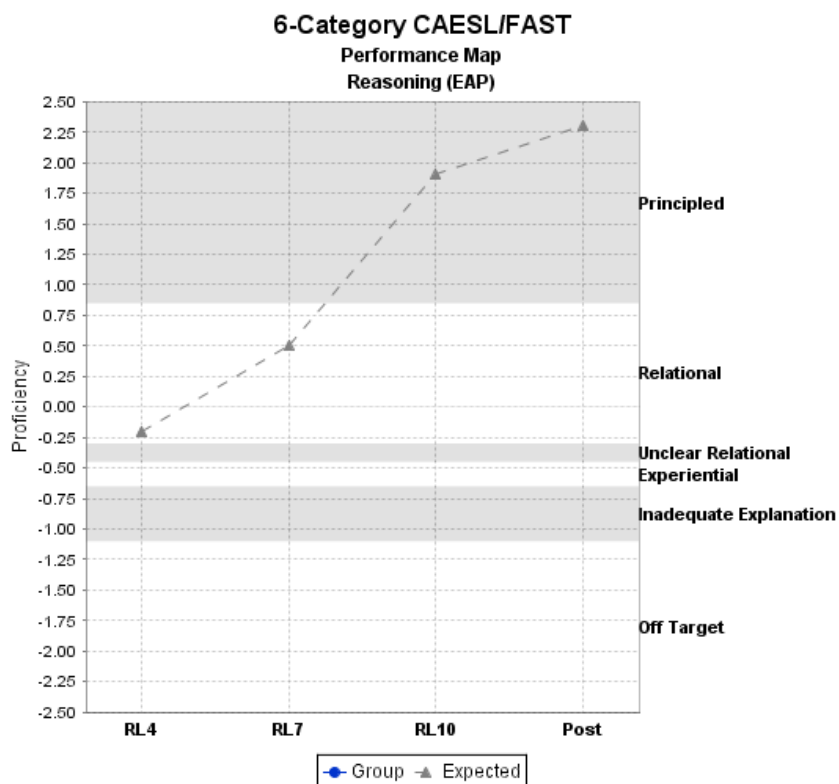


Figure 8. *Performance Map showing progression expected by the curriculum for the Reasoning variable of the CAESL/FAST curriculum for the six-category model.*

Three-Category Model

The underlying theory of a developmental progression from misconceptions through an understanding of relative density is the same in the three-category model as in the six-category model. However, the score levels on the three-category model represent this progression differently. Each progress variable is represented by three levels of knowledge. For WTSF knowledge is represented as increasing from a level describing students who have not grasped the targeted concepts, to a level describing students who are actively working to understand the concepts (i.e., they are using the concept correctly at least 50% of the time), to a level describing students who are actively working to understand a more advanced concept and have, at least theoretically, mastered the targeted concept. The WTSF criterion zones were computed individually for each instrument. Table 3 shows the upper and lower limits of the “On Target” zone for each instrument.

Table 3.

Upper and Lower Limits of the “On Target” category of the 3-category WTSF progress variable for the CAESL/FAST curriculum.

Assessment	Lower Limit	Upper Limit	Targeted Knowledge
RL4	-2.5	0.1	Mass or volume
RL7	-2.0	-0.6	Mass and volume
RL10	-0.5	0.5	Density
Post test	-0.25	0.3	Density

For the Reasoning variable the levels define a progression from needing help, to using well-defined relationships, to using general principles. The Reasoning variable had a consistently defined meaning for the “On Target” level across all assessments (i.e., the Relationship level). Figures 9 and 10 show the GradeMap *Construct Maps* for the WTSF and Reasoning progress variables for the three-category model.

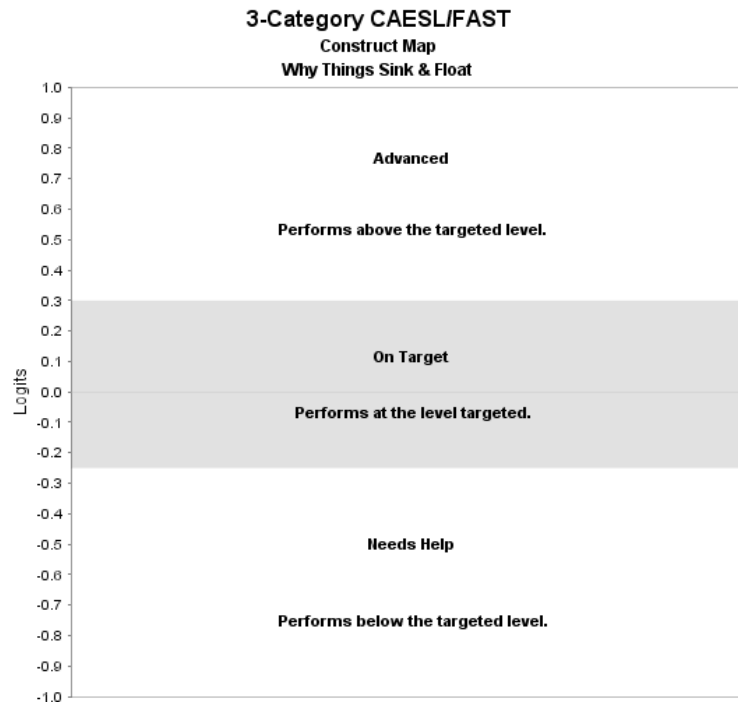


Figure 9. *Construct Map for the three-category WTSF progress variable on the post test of the CAESL/FAST curriculum.*

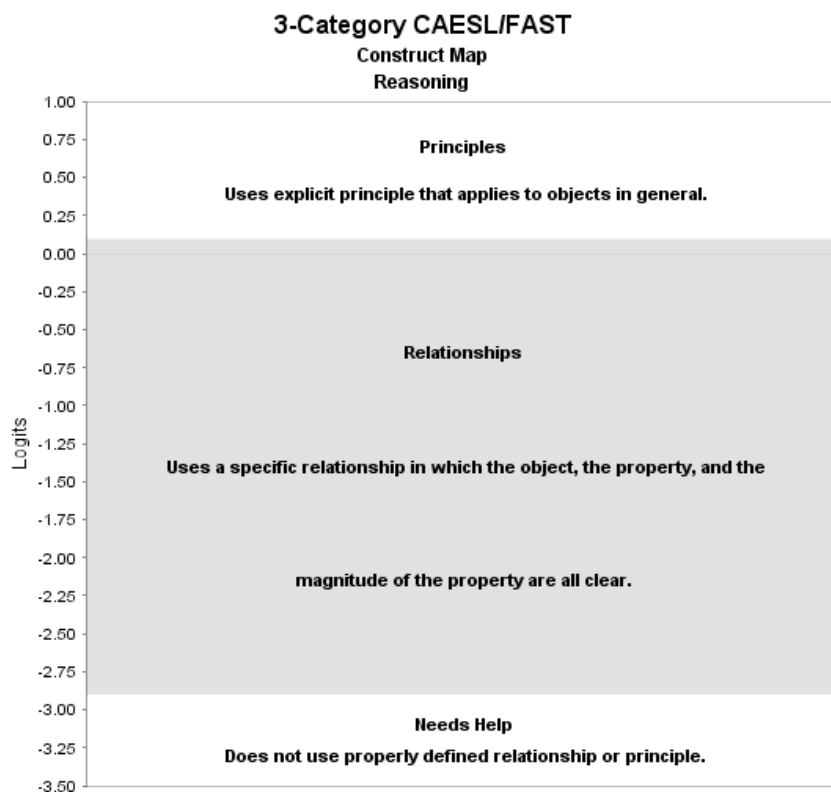


Figure 10. *Construct Map for the three-category Reasoning progress variable of the CAESL/FAST curriculum.*

The expected proficiency levels for the embedded assessment instruments (RL4, RL7 and RL10) on the WTSF variable are defined by the midpoints of the “On Target” criterion zones established for each instrument. This straightforward approach could not be applied to the post test, however, because the post test items targeted a variety of levels. On average, the post test items targeted Density concepts, while the instruction preceding the post test targeted Relative Density concepts. For this reason, and to maintain consistency with the learning expectations of the six-category model, the expected proficiency level for the post test was computed from the means of the “Advanced” thresholds.

Similarly, the Reasoning items all target using Relationships, but we hypothesized that the curriculum expects performance at the General Principles level at RL10 and at the post test. So, the expected proficiency levels for RL10 and the post test on the Relationships variable were computed from the means of the “Advanced” thresholds for those instruments. The expected proficiency levels for both variables are shown in Table 4. The expected learning progressions are illustrated in Figures 11 and 12.

Table 4.

Alignment of assessments to targeted three-category response levels on the WTSF and Reasoning progress variable of the CAESL/FAST curriculum.

Instrument	WTSF level		Reasoning level	
RL4	On Target	-1.2	Relationships (On Target)	0.30
RL7	On Target	-1.3	Relationships (On Target)	0.05
RL10	On Target	0.0	General Principles (Adv.)	1.00
Post test	Advanced	1.6	General Principles (Adv.)	1.50

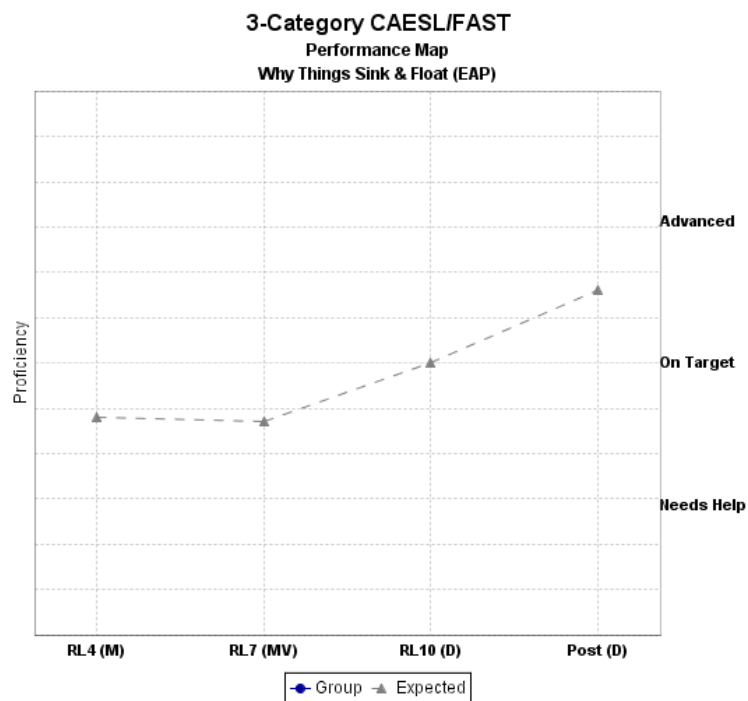


Figure 11. Performance Map showing progression expected by the curriculum for the three-category WTSF variable of the CAESL/FAST curriculum.

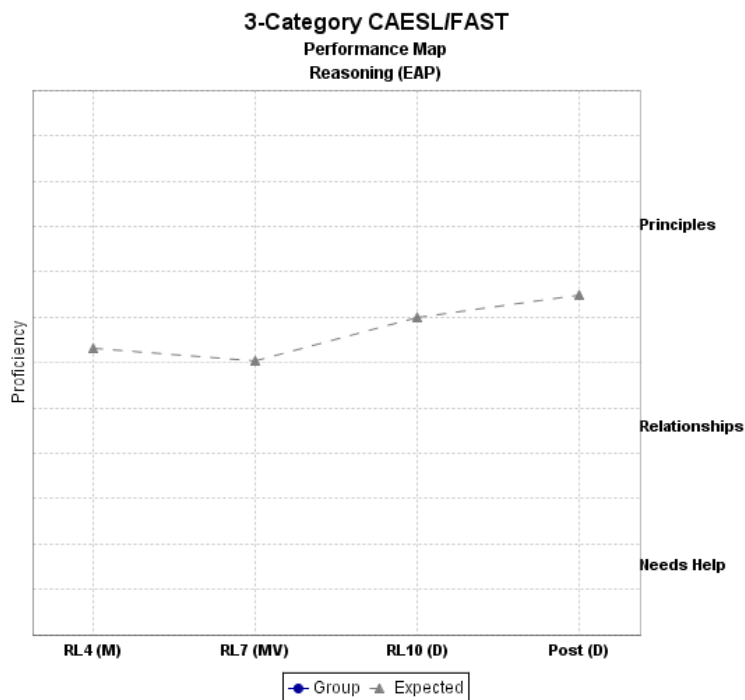


Figure 12. Performance Map showing progression expected by the curriculum for the three-category Reasoning variable of the CAESL/FAST curriculum.

The criterion zones and expected proficiency levels for each scoring approach establish an interpretive context with which student proficiency estimates can be analyzed and compared. Without a context such as the one defined here, the meanings of proficiency estimates are limited to numerical comparisons. Our approach is to construct a qualitative learning progression, based on a theory of knowledge development and operationalized through item content, that describes what students know and can do at various locations along that progression.

Findings

Analysis of Post Test Instruments for Violations of IRT Assumptions

In this part of the study, we examine the psychometric properties of the post test instrument under both models. In particular, we are trying to determine if both models uphold the basic assumptions of IRT modeling: each subscale is unidimensional, higher scores on an item are associated with higher overall ability estimates, and the items within each subscale are conditionally independent.

Dimensionality - We asked three questions about our theorized dimensionality of the model: Do we need additional dimensions to model the data? Could we use fewer dimensions to model the data? and Is there an advantage to modeling the dimensions together rather than separately? To answer the first question, we examined the mean square fit statistics (Wright and Masters, 1981) of the items within each dimension. This statistic is a ratio of the observed variance in the data to the variance expected by the model; when the value is near 1, the observations are measuring as intended. When they are significantly larger than 1, then something other than the intended latent

variable may be captured in the observations. When significantly smaller than 1, the observations are exhibiting less stochastic variability than we expect for good measurement.

We found that the items within each subscale operated in a consistent manner for both the six- and three-category models. None of the item mean square fit statistics for WTSF or Reasoning fell outside the range of 0.75 and 1.33 for the six-category model, and only one item fell outside that range for the three-category model. That one item, on the WTSF variable, had an infit mean square value of 0.70, indicating that responses to this item exhibited less variance than we would expect. The overfit of this single item should not have a particularly deleterious affect on the proficiency estimates produced by the instrument overall.

To answer the second question, we constructed a unidimensional comparative model that treated all of the responses as indicators of a single latent variable and compared the overall fit of that model with our theorized model using a χ^2 test; the unidimensional model is a constrained, or nested, version of the two-dimensional model.

The two-dimensional models provided a better fit to the data than the unidimensional models for both the six-category and three-category approaches. In both cases, the two-dimensional models produced smaller -2 log likelihood statistics than the unidimensional models. The difference in the statistics was statistically significant, indicating that the two-dimensional models provided significant improvement over the unidimensional models in fitting the data (a difference of 459 was obtained for the six-category model and a difference of 43 for the three-category model, both with 2 degrees of freedom where $\chi^2 = 5.991$ at the $\alpha = .05$ level).

To answer the third question, we constructed additional unidimensional models that ignored responses on the Reasoning variable, calibrated that model, and computed proficiency estimates from the posterior distribution based on the WTSF items only. These estimates were compared to proficiency estimates from the two-dimensional models in which the covariance of the two dimensions was part of the equation. Standard errors, person fit statistics, and IRT person reliability indices were compared to determine which estimates were more reliable representations of person proficiency. The two-dimensional models provided more reliable EAP estimates of WTSF proficiency than the models comprised of only WTSF items, for both the six-category and three-category approaches. When we calibrated the two-dimensional models, we found high correlations between the variables (0.941 for the six-category model and 0.924 for the three-category model), indicating that knowledge about WTSF and the ability to explain WTSF using general principles are highly correlated. Charts of the relationship of standard errors of measurement to proficiency estimates are shown in Figure 13. The figure illustrates that the two-dimensional models produced less error in the estimates than the unidimensional models. In the figure, “Complex” refers to the two-dimensional model, while “Simple” refers to the unidimensional model.

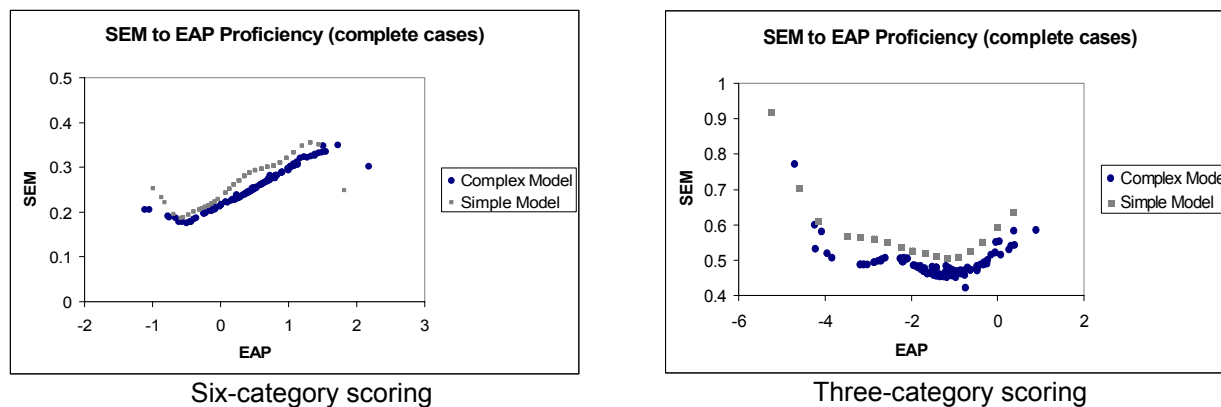


Figure 13. *Standard errors compared to EAP estimates for the six- and three-category scoring approaches.*

Paired-sample t-tests were also conducted to evaluate the significance of the differences in the EAP estimates and associated standard errors of measurement. The mean differences in the proficiency estimates and in the standard errors were significant at the $\alpha < .0005$ level; the two-dimensional proficiency estimates were larger than the estimates under the unidimensional models, while the standard errors of measurement for the two-dimensional models were smaller than the errors for the unidimensional models.

A final comparison was conducted of reliability for the two models using IRT-based person separation indices. Person separation is an index of the consistency of the person measures (i.e., if the instrument were administered again) computed as a ratio of the variance in “true” person abilities over the variance in the observed abilities. Both the six-category and three-category two-dimensional models produced higher reliability indices for the EAP estimates of WTFSF proficiency (.82 compared to .77 for six-category scoring, and .77 compared to .72 for three-category scoring).

Our conclusion from these analyses is that for the purpose of obtaining reliable estimates of WTFSF proficiency for formative assessment purposes, the theorized two-dimensional models provided a better fit to the data than the alternative unidimensional models we tested.

Monotonicity - To examine the assumption of monotonicity, we analyzed the mean proficiency estimates for each category of an item. This analysis is similar to evaluation of point biserials in classical test theory (CTT), except that we use IRT proficiency estimates rather than raw scores. When the assumption holds, higher proficiency estimates are associated with higher scores on individual items. Although we found quite a few items (16 out of 24) in which the mean proficiency estimates of adjacent categories were mis-ordered in the six-category model, only one of those differences was statistically significant at the $\alpha = .05$ level (e.g., the 95% confidence intervals for the estimates in adjacent scoring categories overlapped significantly), and that appears to be due to a single respondent with a high proficiency who obtained a score of 0 on the item. We did not observe this problem with the three-category scoring model.

The left panel of Figure 14 shows an example of the mis-ordering that was common among the six-category items; 11 cases involved only the two lowest score levels and five involved only one respondent in the mis-ordered category. The right panel shows the item for the three-

category model. The vertical bars show the range of the proficiency confidence intervals for each category, with the mean of the intervals designated by the short horizontal marks. The x-axis displays the response values along with the percent of respondents whose responses were scored at those values.

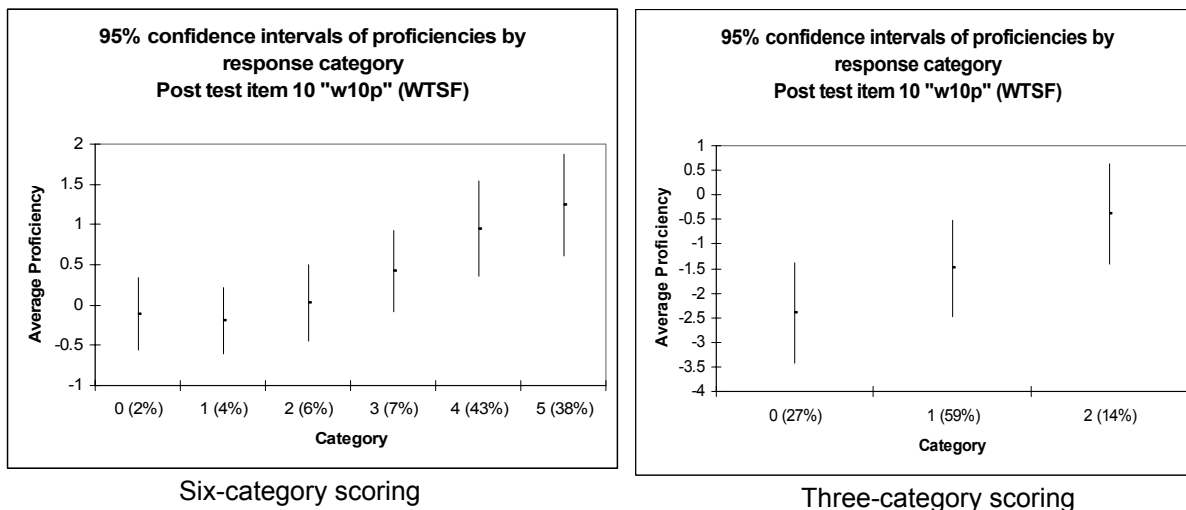


Figure 14. 95% confidence intervals of average EAP estimates for each scoring category for item "w10p" (item 10, post test, WTSF score) for the six-category and three-category scoring approaches..

Local Independence - Violations of local independence typically occur when the response to one item depends upon the response to another item. This is common in assessments where one stimulus is associated with multiple responses that are scored on the same latent variable. For example, selecting an equation, computing the result, and answering the original question may each generate a dichotomous score on a mathematics latent variable. However, either selecting the incorrect equation or making a computational mistake is likely to cause the final answer to be incorrect. In this case, we would say that the third response is dependent upon the quality of the first and second responses.

Violations of local independence within a subscale were not considered likely in the post test. Each prompt elicited a single response, which was scored on each variable. Although conditional dependencies are likely between scores on the same item, this association between latent variables is already modeled in the covariance matrix. No causal relationships between items within a dimension appeared likely in the post test because none of the item responses depended upon responses from other items.

From these analyses, we conclude that both the six-category and three-category two-dimensional models provide a sufficiently good fit to the data so that interpretation of the proficiency estimates produced by the models will be reasonable. At this point, we turn from the psychometric properties of the two models to using the assessment data produced by each model to construct assessment stories about students.

Assessment stories about groups of students

When a teacher is interested in seeing how students are responding to assessment items relative to the goals of the curriculum, he or she may review a *Frequency Map*, which displays the distribution of proficiency estimates for all students on each of the progress variables. Figure 15 shows *Frequency Maps* of the proficiency estimates of the WTSF progress variable for 39 students from one class in the study at four time points: at RL4, at RL7, at RL10, and at the post test using the six-category scoring model.

The upper left chart shows performance at the first time point, the upper right is at the second time point, the lower left is the third time point, and the lower right chart shows estimated proficiencies at the end of the unit. The headings across the top of each chart, and the associated vertical shading, indicate the criterion zones for the six-category WTSF progress variable; starting from less sophisticated understanding on the left (denoted as “OT” for Off Target) to the highest expected level on the right (denoted as “RD” for Relative Density). The percentages of scores are shown along the y-axis, while the proficiency levels are shown along the x-axis.

Teachers can also obtain a report of the names of students with proficiency estimates in each category and the exact percentage in each category from the *Abilities by Level* report. For example, at the post test, 4% (1 student) were at the Mass or Volume level, 82% (23 students) were at the Mass and Volume level, 14% (4 students) were at the Density level, and no students were performing at the Relative Density level. The class chosen to illustrate overall student performance is not representative of students in the pilot study, but is particularly useful as an example of identifying students who are in need of additional support. The progression illustrated in these charts shows that most students did progress over time, but not to the extent anticipated by the curriculum.

The assessment story told from this sequence of charts is that at RL4, most students were at least using mass to explain floating and sinking, and some students were starting to use the relationship of mass to volume in their explanations. At that point in time, students seemed to be performing about as expected. At RL7, when students were expected to be using the relationship of mass to volume in their explanations, fewer than half of the students were doing so. Some students appear to have become confused and were giving responses that exhibited misunderstandings of the basic properties of mass and volume and their influence on buoyancy. At RL10, those misunderstandings appear to have been resolved and students were using the relationship of mass to volume to explain buoyancy. However, few students were using the concept of density to explain buoyancy, despite just completing lessons that focused on that concept. At the end of the unit, most students were still using the relationship of mass to volume to explain buoyancy.

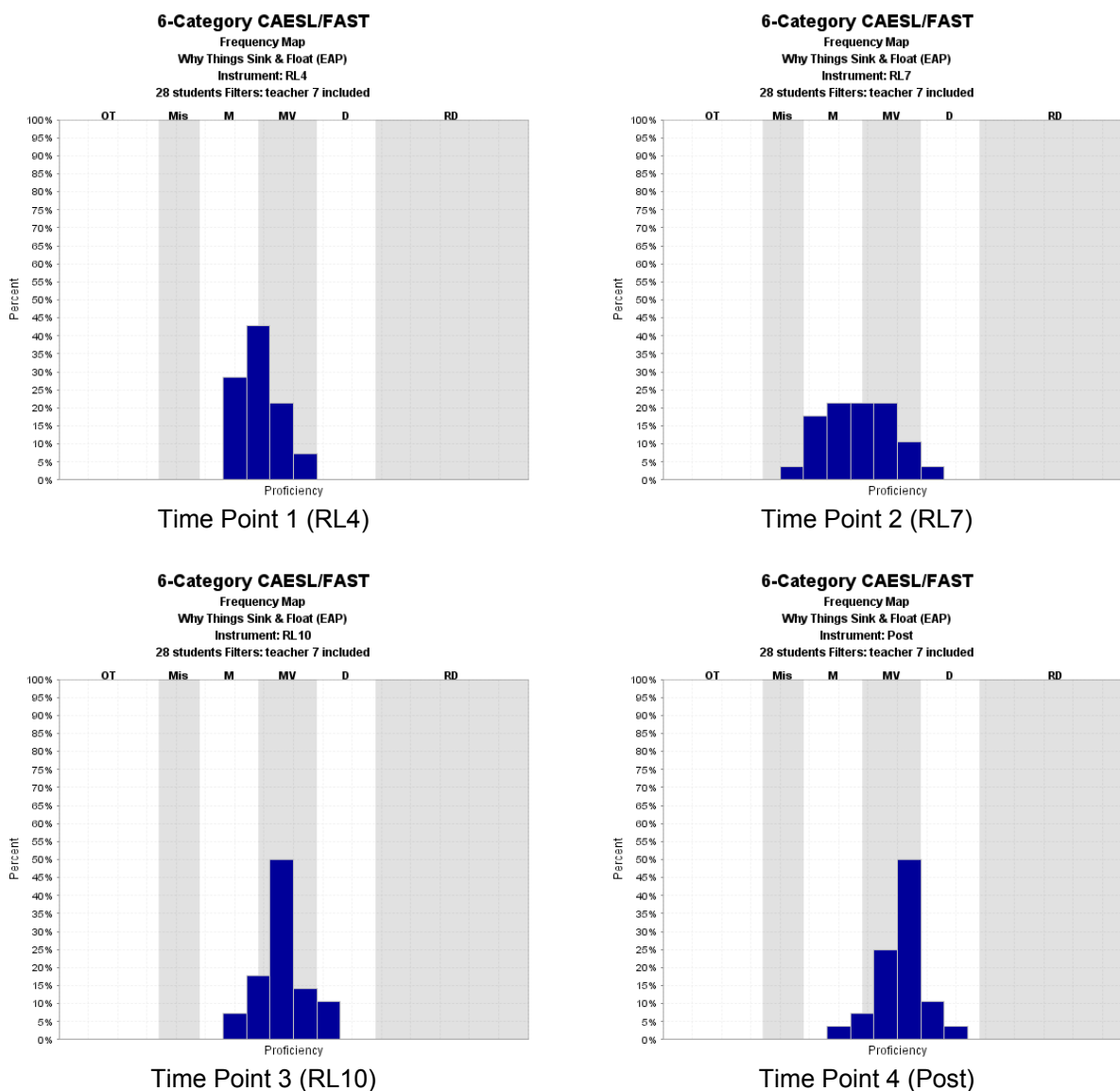


Figure 15. Frequency charts for the *Why Things Sink and Float* variable at four time-points in the CAESL/FAST curriculum for one class ($N=28$), six-category model.

A teacher would normally view these charts immediately after scoring student work for one assessment to determine next steps in instruction for the class as a whole. He or she might feel that at RL4 students were doing pretty well and do not need review before proceeding to the next lesson. At RL7, he or she may want to work with the whole class to uncover the nature of the misunderstandings students are exhibiting in their responses. The problem may simply be the way students are explaining their answers, rather than their actual level of knowledge. At RL10, the teacher may want to review the concept of density and its meaning, since students seem to understand the concept of the relationship of mass to volume fairly well.

Simplified Scoring

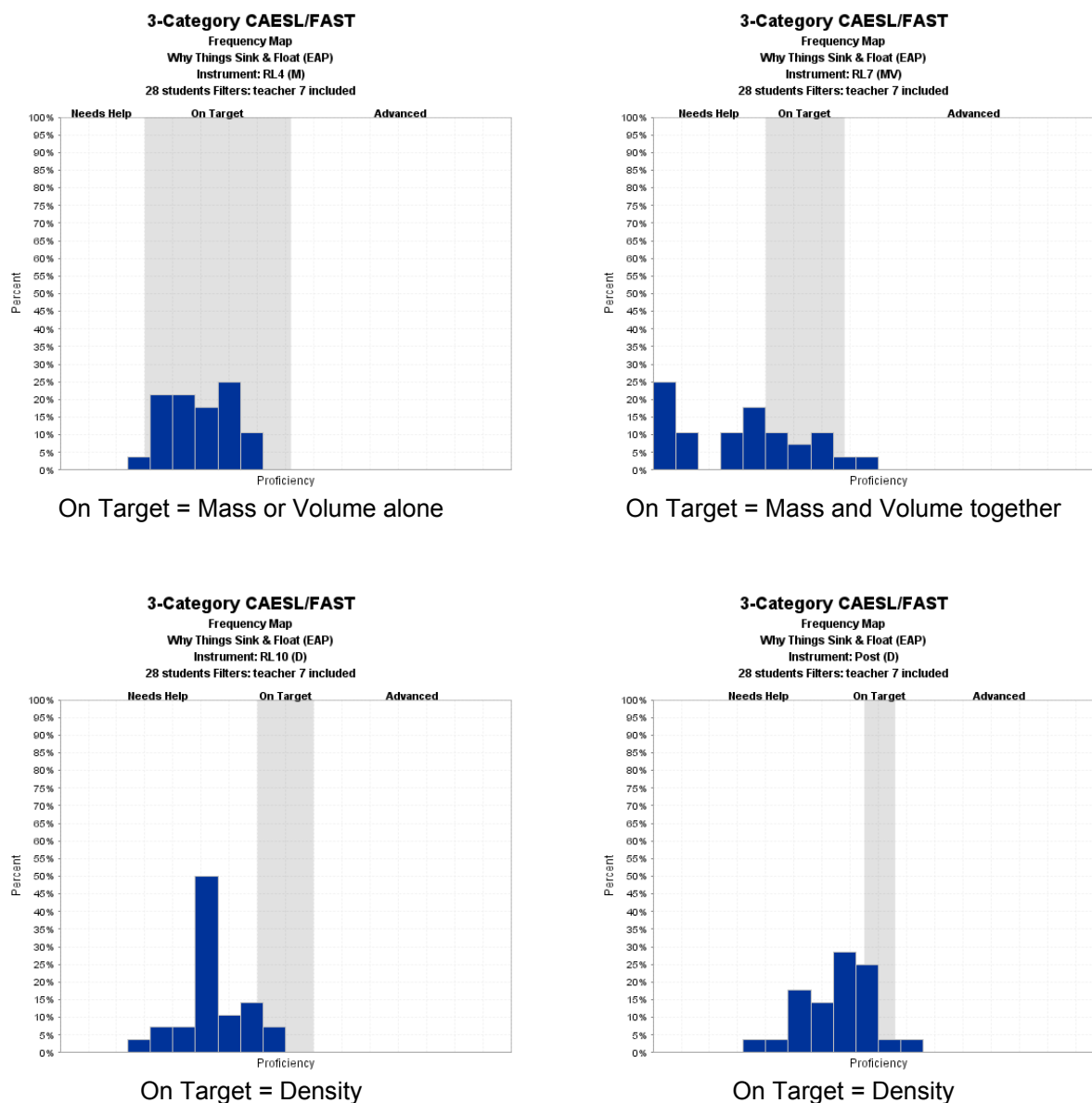


Figure 16. Frequency charts for the three-category *Why Things Sink and Float* variable at four time-points in the CAESL/FAST curriculum for one class ($N=28$).

Figure 16 shows frequencies of the WTSF proficiency estimates at the same four time points using the three-category scoring model. Interpretation of these reports is not as straightforward as we found with the six-category model, because the categories have different interpretations for each instrument and time point. Note that the targeted level of each instrument is noted in the instrument name, located in the heading of each chart. For the first time point, “On Target” means that the students in that category are doing about as expected in understanding how mass or volume alone affect floating and sinking. At the second time point it means that students are

understanding how mass and volume work together to influence floating and sinking, and at RL10 and the post test it means students are using density to explain floating and sinking.

In these reports, aggregated student performance is more stringently aligned with the expectations of the curriculum than in the six-category model. Nuanced and productive understandings of targeted concepts are simply relegated to the lowest score level, and highly advanced understanding falls into the same category as slightly advanced understanding.

Using only the data displayed in these *Frequency Maps*, we would say that students in the class were struggling with the concepts targeted by the curriculum. At RL7 the majority of students were not yet using the relationship of mass to volume to explain buoyancy, and at RL10 very few were using density even 50% of the time in their explanations. If we also review the *Abilities by Level* report, we would find that 4% (1 student) were more advanced than the level targeted by the post test (the Density level), 14% (4 students) performed at the targeted level, and the vast majority, 82% (23 students), were performing below the targeted level at the post test.

Assessment stories about individual students

The expected proficiency levels targeted by each assessment time point provide a useful criterion-referenced progress benchmark for evaluating student performance at that point in time. In addition, we can examine student progress over time to help ensure that students are progressing even when they do not meet specific curricular expectations. Figure 17 shows overall progress through the curriculum for one student using the six-category WTSF scoring guide. We remind the reader that the proficiency estimates reported by GradeMap are not competency measures, but rather an indication of what a respondent is about to understand more fully. When a proficiency estimate for a student lands in a particular criterion zone, it means the student is actively using the concepts associated with that zone about half of the time.

Each point shows the level of understanding the student is exhibiting (about 50% of the time) after completing one instrument. Thus, these points tell us about concepts students are currently in the process of assimilating into their mental model, rather than concepts that have been mastered. Some *targeted support* at this time could be highly beneficial to the student firming up his or her understanding of the concept.

Student 129, displayed in Figure 17, was most actively engaged in learning the concept of the relationship of mass to volume at RL4. This implies that he had mastered the concept of mass influencing floating and sinking by that time. At RL7, he appears to have mastered the concept of the relationship of mass to volume and is using the concept of density to explain why things sink or float. At RL10, he has developed more ability, but is still learning the concept of density. At the post test, he is in the process of learning the concept of relative density and probably understands the role of density in buoyancy more completely.

The three-category model, shown in Figure 18, provides a similar story about the student. At RL4 he is performing at or above the mass level, at RL7 he is at or above the mass and volume level, at RL10 he is at or above the density level, and at the post test he is performing at the relative density level. Note that this chart does not show shaded criterion zones. This is because

the zones differ for each instrument. The labels on the right edge of the chart are a general guideline, but do not have specific limits.

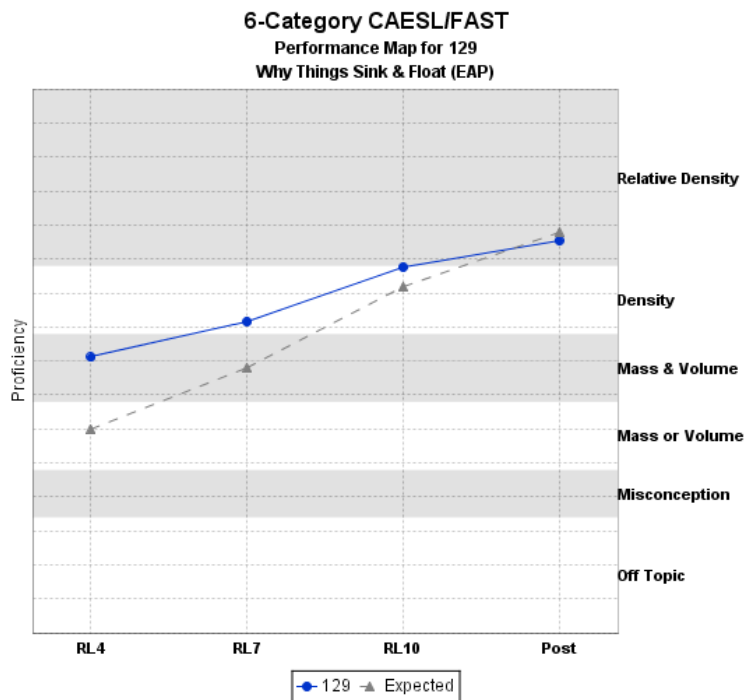


Figure 17. Map of progress for Student 129 on the six-category WTSF variable for the CAESL/FAST curriculum.

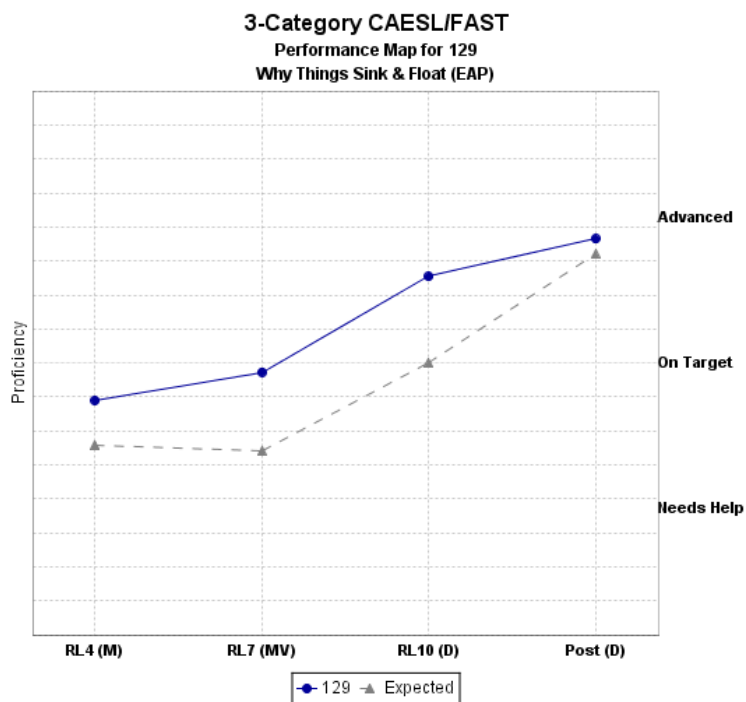


Figure 18. Map of progress for Student 129 on the three-category WTSF variable for the CAESL/FAST curriculum.

Student 822 has a different experience with the curriculum, as shown in Figures 19 and 20. She begins with a general understanding of how mass affects an object's capacity for floating or

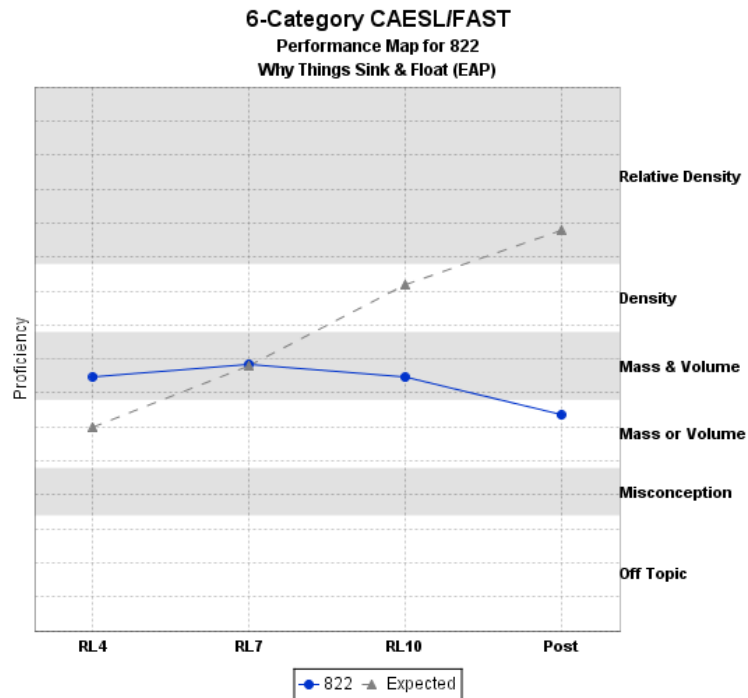


Figure 19. Map of progress for Student 822 on the six-category WTSF variable for the CAESL/FAST curriculum.

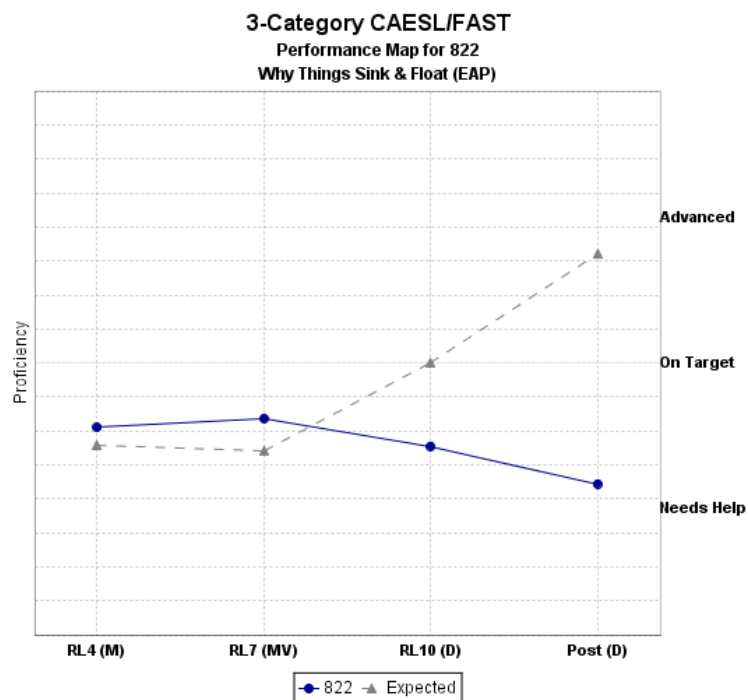


Figure 20. Map of progress for Student 822 on the three-category WTSF variable for the CAESL/FAST curriculum.

sinking, and is exhibiting some understanding of the relationship of mass to volume at RL4 and at RL7. However, by RL10, she is still at the same level of understanding, and by the post test, her responses indicate that she has not fully understood how mass or volume alone influence floating and sinking. The three-category model, illustrated in Figure 20 provides similar information to the six-category model illustrated in Figure 19.

Of course, teachers would not normally wait until the unit is completed before looking at student progress. Normally, when the student stops showing progress, at RL10 in this example, the teacher would think about providing some kind of support to the student. The teacher may find it useful to know how the student performed on each item to gain a better understanding of the problem she seems to be having. A GradeMap *Diagnostic Map* can provide more detailed information.

The *Diagnostic Map* from the six-category model for Student 822 for the RL10 assessment activity is shown in Figure 21. This report provides a wealth of information to a teacher. The data are arranged in three columns. The left column shows the response levels that student achieved on each item of the assessment activity. The items are ordered from the least difficult item at the bottom, in this case item “RL10d,” to the most difficult item at the top, item “RL10b.” The score the student earned is displayed after the “.” next to the item name. Student 822 responded at the “Misconceptions” (Mis) level on item RL10d, and at the “Density” (D) level on item RL10b. The center column displays the location of the student’s proficiency estimate after completing the assessment with the “XXX” indicator. Since proficiency estimates include an error of measurement, the horizontal dashed lines above and below the proficiency location indicate the most likely range of proficiency that the student is exhibiting, plus or minus one standard error from the computed estimate.

All of the items displayed in this range, including those at the border, are the responses the student is expected to make about 50% of the time. In this example, we expect Student 822 to respond to item “RL10b” at the “Density” level about 50% of the time. In fact, we know that the student did respond at that level on this assessment because the item is listed in the left column. On the other hand, we also expected the student to respond to item “RL10c” at the “Mass and Volume” level and to item “RL10d” at the “Mass” level about 50% of the time, but she did not. Items that appear in the right column indicate a response level one category higher than what the student actually achieved (i.e., a score at the “Mass” level on “RL10b” and a score at the “Misconceptions” level on “RL10d”). We are not surprised by this result, because we only expected the student to achieve those scores about 50% of the time.

Item responses that appear below the range of expected proficiency for the student (below the lower dashed line) are those that we expect the student to achieve because they are less difficult than the student’s proficiency implies that he or she can perform. In this example, we expected the student to achieve a score at the “Mass and Volume” level on item “RL10at” and scores at the “Mass” level on items “RL10c” and “RL10ae.” Item responses that appear above the range are those that we would not expect the student to achieve because they are more difficult than we expect the student to be able to perform. In this example, we did not expect the student to achieve a score at the “Density” level on item RL10at (in the right column) or a score at the

item “RL10b”. The implication is that the student obtained the lowest score, “-“ (needs help) on all the other items. Since the general level of knowledge targeted by RL10 is “Density,” the teacher would know from this report that the student is having difficulty with that concept. The placement of item “RL10d” with an “x” response in the right column, below the range of expected student proficiency denoted by the dashed line, suggests that the student should have been able to respond with an answer involving density. Review of this item with the student may help uncover a nagging misperception the student is dealing with. Items “RL10c,” “RL10ae,” and “RL10at” at the “x” level, which all appear above the student’s proficiency range, suggest that the student is still working to understand the concept of how mass and volume work together, and is not ready to focus on the concept of density.

Although the diagnostic information is not as detailed with the three-category model as it was with the six-category model, both *Diagnostic Maps* for Student 822 suggest that focusing on the concept of how mass and volume together relate to an object floating or sinking would be most useful for this student at this point in time.

Once the criterion zones for each model were defined, the assessment stories emanating from the three-category scoring approach were similar to those from the six-category approach when they dealt with groups of students. For individual students, the assessment stories from the six-category approach were much more specific than those from the three-category approach in identifying instructional next steps. Ongoing research focuses on improving the diagnostic value of a three-category approach.

Conclusions

Curriculum developers, assessment designers and psychometricians are discovering new and better ways of evaluating complex, performance-based activities and drawing useful inferences about student knowledge, abilities and skills from such evidence. If we want teachers to embrace these developments in assessment capabilities, then we need to attend to the real constraints teachers experience: constraints of both time and energy to learn something new.

This paper reported some of the detailed information that can be obtained by applying complex scoring rubrics to the evaluation of classroom activities, but the demand on teachers’ time to score and interpret results using those rubrics is considerable. The findings of this study suggest that scoring student work using a simpler generic three-category rubric with score levels of “advanced,” “on target,” and “needs help” can reliably reproduce some of that information with less effort required on the part of the teacher. The three-category model we implemented was particularly useful for identifying students who were not performing at the targeted level and in providing information regarding the concepts those students would benefit from focusing on right away. The approach was less useful for providing the diagnostic detail teachers might like to have about all students in a class.

The calibration and alignment of a simplified scoring model that can be used to evaluate student progress over time, as opposed to evaluating knowledge at a single point in time, involves a number of steps to establish criterion zones that define qualitatively distinct levels of student knowledge. We were able to identify individual criterion zones for each instrument that categorized students in a manner consistent with the categorizations obtained with the six-

category scoring approach. A logical next step in this research is to explore transformation processes to align the instrument-specific three-category criterion zones with the six-category rubrics. This type of alignment would potentially provide as much information from the three-category scoring as is currently obtained from scoring with the six-category rubrics.

Of course, such an approach is most appropriate for curricula that have well-defined developmental perspectives of learning, where specific levels of performance can be articulated and in which student progress means traversing through lower levels to arrive at higher levels. In addition, the use of software to transform scores into visual representations of student knowledge is particularly helpful when the observed data are complex.

Once the psychometric details have been worked out, the approach will be tested with several curricular units in classrooms to obtain feedback from teachers about the viability of the three-category scoring approach. Research will focus on the usefulness of the formative feedback provided by the GradeMap software and the usability of the software as a formative assessment tool.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 2*(1), 1-23.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47-76.
- Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5).
- 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.
- Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2005). *The nature and impact of teachers formative assessment practices*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 2005.
- Howe, A. A. (1997). Reliability study on the use of a rubric in elementary science. : Adams State College. Alamoso, Colorado.
- Kennedy, C. A. (2005). *The BEAR assessment system: A brief summary for the classroom context* (No. 2005-03-01). Berkeley, CA: University of California, Berkeley Evaluation & Assessment Research Center.
- Kennedy, C. A., Brown, N. J. S., Draney, K., & Wilson, M. (2005). *Using progress variables and embedded assessment to improve teaching and learning*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 2005.

- Kennedy, C. A., Wilson, M., & Draney, K. (2005). *GradeMap v4.2*. Berkeley, CA: University of California, Berkeley Evaluation & Assessment Research Center.
- National Research Council. (2001a). *Classroom assessment and the national science education standards*. Washington, DC: National Academy Press.
- National Research Council. (2001b). *Knowing what students know*. Washington, D.C.: National Academy 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.

Reference as:

Kennedy, C.A. (2006, April). *Simplified scoring of performance activities: Comparing assessment stories from complex and simple scoring approaches*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, April 2006.