Improving Assessment Evidence in e-Learning Products: Some Solutions for Reliability

Kathleen Scalise[1*], Tara Madhyastha[2], Jim Minstrell[2], Mark Wilson[3]

*[1]Assistant Professor, Department of Educational Methodology, Policy and Leadership, University of Oregon, Eugene, OR 97403, USA, **kscalise@uoregon.edu***
[2]Research Scientist, Facet Innovations, LLC, 1314 N.E. 43rd St., Suite 207, Seattle, WA 98105, USA, former affiliation, currently reachable at tara.madhyastha@gmail.com
[3]Senior Research Scientist, Facet Innovations, LLC, 1314 N.E. 43rd St., Suite 207, Seattle, WA 98105, USA, jimminstrell@facetinnovations.com
[3]Professor, Policy, Organization, Measurement and Evaluation, University of California, Berkeley, CA 94720, USA, MarkW@berkeley.edu
**corresponding e-mail: kscalise@uoregon.edu*

Abstract:
E-learning products such as cognitive diagnosers interact with learners and collect assessment data to build a picture of some aspect of a learner's thinking. One concern for this rapidly emerging area of e-learning is whether the diagnostic conclusions of such products are based on sound evidence, including whether or not the diagnostics are reliable. In online settings, the information may be used for adaptive delivery of content, individualizing learning materials, dynamic feedback, teacher feed-forward, cognitive mapping, score reporting and course placement. A reliability index quantifies the impact that measurement error at the individual level may have on the accuracy of the inference. This paper investigates some simple solutions that substantially improve reliability within one e-learning product. These solutions include providing questions of appropriate difficulty that help to maximize item information across the distribution.

Keywords: assessment, e-learning, cognitive diagnoser, reliability, adaptive delivery, personalized, online, feedback, measurement, diagnostic, physics education.

E-learning products with assessments such as cognitive diagnosers interact with learners and collect assessment data to build a picture of some aspect of a learner's thinking (Dimitrova, 2002; Kennedy, Bernbaum, Timms, Harrell, Burmester, Scalise, & Wilson, 2007; Scalise, Bernbaum, Timms, Harrell, Burmester, Kennedy, & Wilson, 2006). Such diagnostics are rapidly being included in e-learning products. A common purpose is to adapt the flow of materials so that each student receives content that is personalized to meet particular needs (Scalise & Claesgens, 2005; Taylor, 2002; Trivantis, 2005; Turker, Görgün, & Conlan, 2006). Diagnostic approaches may be used to provide information to help teachers and students understand learning performance, make appropriate course placements, and differentiate instruction in the classroom.

One motivation for differentiated instruction (Tomlinson & McTighe, 2006) is that traditional curricular materials and assessments sometimes are not helpful for where the student is in the knowledge acquisition cycle (Gifford, 1999; Hopkins, 2004). By comparison, differentiated instruction approaches are seen as moving teaching and learning activities toward the needs of the student. In practice, differentiated instruction that depends on individual learning evidence can be difficult to implement in a classroom environment, where teachers have little time to address individual needs.

Technology can help instructors lower the resource barrier for this type of differentiated instruction. E-learning products can also marry potentially powerful assessment tools with new information technologies to capture and analyze student data, rapidly deploy new media, facilitate collaboration, and provide other e-learning amenities such as asynchronous learning (Gifford, 2001; Parshall, Davey, & Pashley, 2000). In online settings, the information collected about the needs of different students may be

used for adaptive delivery of content, individualizing learning materials, dynamic feedback, teacher feed-forward, cognitive mapping, score reporting and course placement (Gifford, 2001).

Technology to deliver differentiated instruction is now readily available, with back-end databases and a variety of multimedia-rich streaming techniques for which the flow of content to students can be adjusted in near real-time (Turker, Görgün, & Conlan, 2006). Reports of student performance can also readily be generated to provide information for instructors and others. However, the usual measurement concerns of high quality data and inferences can quickly derail efforts to make inferences in an accurate and speedy fashion (Osterlind, 1998; Wilson & Scalise, 2003), threatening to undermine the usefulness of dynamically personalized learning objects and diagnosers.

A major concern with personalized e-learning and assessment is whether the diagnostic conclusions of such products are based on sound measurement evidence, including whether or not the diagnostics are reliable (AERA, APA, & NCME, 1999; Pellegrino, 2005; Pellegrino, Baxter, & Glaser, 1999). Reliability has to do with the consistency of measurement, and whether results can be reproduced. A reliability index can help quantify the impact that measurement error at the individual level may have on the accuracy of the inference (Adams, 2006; AERA, APA, & NCME, 1999). Low test reliability suggests that less confidence can be placed in the diagnosis.

An e-learning example

This paper investigates reliability evidence for one cognitive diagnostic product that illustrates some of the reliability challenges in e-learning products. After describing

these challenges, we demonstrate some simple solutions for improving reliability with only minimal increases in the number of tasks and questions that need to be completed.

Like many classroom-based assessments, e-diagnosers are often developed to assess how a student is performing. One goal of cognitive diagnosis often is to determine whether students have met learning objectives. If students have not met objectives, the diagnostic product may attempt to identify the reasons, and what partially correct or productive ideas students might have on which one might base further instruction.

This is a complex set of goals and presents some substantial challenges to achieving reliable evidence. First, it is often important to accurately measure student thinking within limited student time, using relatively few assessment questions and tasks. Practically, students can't be asked to respond to too many questions or it will overwhelm them or the teachers, who may have to score and interpret the outcomes. Teachers, curriculum developers and others working on instructional design legitimately want to ensure that most of the available instructional time is devoted to learning, and assessment can sometimes be seen as peripheral to the learning experience. Issues of reliability, and the associated confidence that can be placed in a student score or measure, are sometimes associated with too little assessment data on each student. But given the time and resources constraints, simply increasing the number of assessment tasks considerably to measure each learning objective is often a difficult option.

Second, the goal of identifying whether a student has mastered the material is sometimes considered more important than identifying other productive incorrect ideas. In this case, questions are often devoted to measuring mastery. If instruction is successful, students may be expected to answer as many as 80 percent or more of the

questions. This may vary, but is similar instructionally to the idea of achieving a B- or better when the percentage correct for achieving such a score may be considered 80 percent and above . Exhibiting successful mastery of numerous assessment items is good in terms of learning. From a measurement perspective, however, more test information is available when students have opportunities to achieve or not achieve an item. Students whose level of ability is well above the task are less well measured for their current state of knowledge. This introduces more uncertainty into the student score, and reduces overall test reliability.

As with all assessment contexts, if low reliability is found in a cognitive diagnoser, it could also be endemic to the goals of what is being measured. The theory behind diagnoser products is that there is sufficient regularity in a learner's reasoning response for an interpretable diagnosis of whatever is being measured. However, students may answer inconsistently across questions and tasks, and this may be a characteristic of students when they are at critical points in the learning process. For example, Tatsuoka and colleagues describe the inconsistency with which students use procedural rules to solve arithmetic problems until they reach mastery and then use the correct rules (Tatsuoka, Birenbaum, & Arnold, 1989). Halloun and Hestenes note that the "common sense" conceptual systems of students are not always internally consistent from the observer's point of view (Halloun & Hestenes, 1985). Such a phenomenon could substantially reduce the reliability of assessing reasoning approaches, but the challenge may be more inherent to the underlying theory than to the measurement instrumentation.

In any case, whenever assessment data includes too much "noise," or randomness in the student answer patterns, this can contribute to less confidence in the student score,

and lower reliability. So the first question to ask when exploring issues of reliability in e-learning products is to understand what the reliability evidence looks like in the given situation. The example in this paper considers reliability evidence from the DIAGNOSER system (Minstrell, Anderson, Minstrell, & Kraus, in preparation) This is a web-based system that is the next generation of the tool described and prototyped by Hunt and Minstrell, which a teacher could use to diagnose student difficulties in science (Hunt & Minstrell, 2004). The system consists of short sets of questions designed to elicit middle-school and high-school student thinking around specific concepts in physics, properties of materials, and human body systems. In this example, we focus primarily on a question set in physics. The content is aligned with two widely used and cited standards documents describing what such students should understand and be able to do in physics—The National Science Education Standards (National Research Council/National Academy of Sciences, 1996) and the Benchmarks for Science Literacy (American Association for the Advancement of Science, 1993).

DIAGNOSER resources and question sets have been developed and tested by teachers and are based on research into the teaching and learning of math and science. Students receive feedback on their thinking as they work through their assignment. Teachers can access reports on student thinking related to the assigned content (Minstrell, Anderson, Minstrell, & Kraus, in preparation).

DIAGNOSER tasks begin with a question related to the objective of interest, such as a graph of the speed and position of a vehicle for a physical science problem set. Then depending on the student's answer to the question, DIAGNOSER follows up with a series of additional questions or probes, in a computer-adaptive approach that is sequential, or

in other words, with the set of potential pathways planned out in advance. An example is shown in Figure 1. DIAGNOSER ultimately assigns a reasoning "facet," or a description of the reasoning pattern, for the student answer on each question or small bundle of questions and probes.

Facets of student thinking in this context are individual pieces or constructions of a few pieces of knowledge and/or strategies of reasoning (Minstrell, 2001). They are generalizations of comments and responses that students say or actions they take in the classroom, for example when a student has a particular idea and expresses it by answering a question a certain way, or making some kind of prediction. Facets are organized around a conceptual idea or a topic, in a "facet cluster." Within a facet cluster, facets are ordered from 0-10 (correct reasoning strategies) to 99 in terms of how "problematic" the reasoning is. The more problematic a facet, the higher the facet number and the more difficulty a student is likely to have with ideas within this facet cluster and related topics.

When used in a classroom environment, the teacher will elicit students' facets of thinking prior to instruction, and then, if those ideas do not conform to established scientific models and theories, the student will be given a chance to test those ideas with a series of experiments, or prescriptive activities. These activities challenge the students' beliefs and help them to move towards the target level of understanding.

Figure 1 shows part of an example DIAGNOSER question set. As shown in the item set flowchart, the task begins for all students with Question 1. A reasoning facet is assigned based on the answer a student selects. There are some cases in which the student response does not yield enough information to assign a facet. Usually in this case,

additional feedback is provided and the student is reassessed. In Figure 1, facet codes for each answer are shown in brackets following the answer.

On completion of the set, only part of which is shown here, each student has facet evidence on nine questions related to the objectives being measured. The learning goals for this question set involve describing motion and determining the speed of an object, for a specific time from the information given. The information may be given in graphs, tables, pictures, or words. Figure 2 shows the alignment of these goals with the National Standards and Benchmarks.

Measurement Models

With assessment embedded in e-learning instructional materials, there is often a shared context over one or more assessment questions. This allows instruction to be effectively situated. But from a measurement perspective, this also introduces dependencies within the group of items used for assessment. This means that a student's score on multiple items may be affected by the group context in which the questions are presented. Organizing groups of items into bundles, or testlets, and treating them with an item response partial credit measurement model such that one score is generated for the set of common items is one way to address these dependency issues.

Item response models express the probability of an occurrence, such as the correct response on an assessment question or task, in terms of estimates of a person's ability and of the difficulty of the question or task. The analysis also generates estimates of how precise a student score is likely to be and therefore the confidence with which it can be interpreted. Item response models can help provide a basis for evaluating how well an assessment instrument is performing by generating validity and reliability evidence

(Wilson, 2005). The item response model used in this paper is the partial credit model (Wright & Masters, 1982), which can be used to score students on small bundles of items that share common stimulus material (Rosenbaum, 1988; Wilson & Adams, 1995).

The partial credit model is the more general of two polytomous Rasch models (Wright & Masters, 1982) commonly expressed according to Equation 1:

$$P(X_{is} = x \mid \theta_s) = \frac{\exp \sum_{j=0}^{x} (\theta_s - \delta_{ij})}{\sum_{r=0}^{m_i} \exp \sum_{j=0}^{r} (\theta_s - \delta_{ij})} \qquad \textbf{(1)}$$

for item i scored x=0,…,m_i, where X_{is} is the score of student s on item i, x represents a given score level, θ_s represents the parameter associated with the person performance ability of student s, r in the denominator represents a summation of terms over the steps, and δ_{ij} represents the parameter associated with the difficulty of the jth step of item i.

Because the partial credit model represents a summation of terms over the steps, the issue of accumulating credit in a sequential scoring model arises. With alternate or misconceived reasoning facets, in using the partial credit model it is assumed that students who are able to establish and explain the correct reasoning can be considered on average to have surpassed naïve reasoning that is in conflict with the correct reasoning. Therefore students have "achieved" the facet, by surpassing it in understanding (assuming good support and fit for the measures). Students therefore are not expected to actually have procedurally compared incorrect and correct answers and worked backwards on each item, explicitly rejecting the incorrect facet, but rather that we posit a

theoretical concept of something like a knowledge state that when achieved can be considered to supersede the incorrect reasoning.

Guessing is of course an additional consideration for this modeling approach. With adaptive testlet structures that only award a score after a sequence of probes surrounding the reasoning of the original answer, it is much more difficult for students to guess correctly on an item. A correct guess would require a sequence of appropriate responses to probes. This is unlikely to occur randomly. It would also require that students are able to resist attractive distractors that are specifically modeled to represent commonly held misconceptions or alternate conceptions. The degree to which each testlet includes enough probes to make guessing sufficiently unlikely is an area that requires future research. In this item set, a correct answer in the revised item set (see section discussing how reliability was improved) required two or more correct responses to probes for nearly each testlet. Of course, qualitative data along with fit statistics and other evidence can help support these assumptions (Wilson, 2005).

Reliability estimates based on student data

In general, test reliability estimates indicate how much of the variance of performance estimates for any given question set is explained by the correlation with hypothetical "true" scores, or hypothetical scores based on multiple takings of parallel forms. In classical approaches, a desirable .8 reliability would indicate about a .9 correlation with the true score.

For the question set partially shown in Figure 1, the EAP/PV^{i} test reliability estimate for a sample of 1,541 students was .47. As a general rule of thumb, a test reliability estimate of .70 is sometimes a minimum acceptable for performance of an

instrument, and .80 or higher is usually preferred. However, if the goal of the diagnostic product is dynamic feedback and if the feedback is made based only on the last student response, it could be reasonable that a reliable aggregated score may not be necessary. This is true because it is often instructionally appropriate to give students feedback on their responses to questions, whether or not another response at another time or in another situation might have been different.

However when higher reliability is necessary, for instance when instructional decisions are made on the basis of individual student performance, we consider alternate approaches, some of which will be discussed here.

The Spearman-Brown prophecy formula can be used to calculate how many questions would be necessary in order to achieve higher reliability for this question set. Using this formula, to achieve a .7 reliability with items like those in this instrument, we would need about three times as many questions, or about 24 questions total. A .8 reliability would require about 36 questions. Given the nature of DIAGNOSER as a quick check for formative assessment, this large number of questions and the time it would take students to complete them is prohibitive as a solution for reliability.

Using Wright Maps and Standard Error Plots for interpreting e-learning assessment evidence

Figure 3 shows a representation called a Wright Map of the analyzed Speed Set data that helps us better understand why the reliability is low for this question set and what else might be done about it besides tripling or quadrupling the number of questions. The Wright Map is empirically generated by the analysis, and is based on the actual student data. The X's in the left column of the figure show a vertical histogram (histogram turned on its side) of student performance for the sample group. The right column is a "scale" of the performance, showing the difficulty of achieving the various reasoning facets, estimated from the data. The "easiest" facets, which are those that the most students were able to achieve or surpass, are at the bottom of the column. The facets are listed by question number followed by the facet achieved in parentheses. Thus, 1(90) at the bottom of the column indicates the threshold difficulty of achieving a 90 facet on question 1, whereas 5(83) indicates the threshold difficulty of achieving an 83 facet on question 5. Since these facets fall at the same level on the Wright map, they are approximately equally difficult to achieve, for these questions.

Some of the facets in the right column show an item number followed by "Correct," as in 7(Correct), just above the 80–90 facet label and the dividing line to the 70 facets. The "Correct" label indicates the estimated threshold difficulty of achieving the correct answer on each item. The fact that the "Correct" threshold difficulty estimates are spread out over a wide range indicates that, based on this data set, the system does not distinguish between giving a more sophisticated response compared to a correct response, but rather that some items are substantially more difficult to achieve than others.

Note that the scale of numbers on the very left of the map, ranging from -1 to 1 in this case, is a logit or "log of the odds" scale. It can be used to compare the logit location of a student and the logit location of an item and determine a probability estimate of whether any particular item and facet will be achieved. A student X that is adjacent to a facet number, such as the lowest X on the map, indicates that students performing at that level would have a 50–50 chance of achieving the facet located next to them. So students at the lowest X, who are the lowest performing students in this data set, would have a 50-50 chance of achieving a correct answer on items 1 and 2, indicated by 1(correct) and 2(correct) near the lowest Xs. These students would have a higher chance of achieving facets below them on the map and a lower chance of achieving facets above them. Thus a student at the lowest X would be very unlikely to achieve a correct answer on Question 3, the most difficult item, which falls at the top of the map. In contrast, students at the top of the map would be quite likely to answer this question correctly. By considering the locations of the students, it is possible to predict what facets, correct and incorrect, they might display.

On the Wright map, the 90s and 80s facets across items show as easiest, as predicted by the facets framework. This is followed in general by facets in the 70s across items, then facets in the 40s and 50s. For two facets, 7(41) and 4(51) flagged in bold on the Wright map, achieving the facet was somewhat easier than predicted. This is shown by the placement of the facet lower in the column as compared to the 40 and the 50 facets for other items in the question set. This sometimes happens when information in the item is provided that gives a clue or scaffolds the student answer. For instance, 7(41) provides a look-up table for reading data that likely reduces cognitive load and directs student

thinking, making it possible for students to be somewhat more successful in achieving a 40-range facet on this question.

So, what does the Wright Map indicate about the question set and possible solutions for reliability? The evidence to notice is that the column of X's on the left is relatively high compared to the distribution of facets on the right. In fact, the performance locations where most of the items and facets measure is in the less proficient 70, 80 and 90 facets. Yet most of the students are achieving the more proficient 40–50 facets, higher up in the distribution, which only a few parts of a few items measure optimally. Considering the standard error plot in Figure 4, the key observation is that as the student performance estimate goes up, so does the standard error, and in the range where most students are performing, above 0 and higher, the standard error is much higher than for students lower in the range. Only about 26 percent of students fall into the portion of the Wright map where most measurement information is available, and thus the standard errors are lowest in that zone of about -1 to -.5 logits as shown by the standard error plot in Figure 4. Nearly three-fourths of the students fall into the upper portion of the Wright map, beyond the capacity of many of the items to measure and to discriminate among performance abilities. This mismatch of the main part of the student distribution to items of sufficient difficulty, and the associated high standard error in the middle and upper range of students substantially impacts the overall test reliability.

To test whether more assessment evidence higher in challenge level reduces standard errors and increases reliability, the example DIAGNOSER set was slightly redesigned. The primary purpose of the redesign was to move more of the measurement information "higher," or in the direction of somewhat more difficult questions, to lower

the standard errors for the majority of the students. This is relatively easy to do in cognitive diagnosers of this type, as the adaptive bundle of items can simply include additional branches of probes, or alternate probes, for higher scoring students.

Generally, DIAGNOSER sets are intended to be completed in approximately 20–30 minutes of class time and include between 6 and 10 questions, most of which in the original set included an initial question and one or more follow-up questions. The follow-up questions are delivered conditionally, or in other words based on student response to the previous question. At least two questions are paired: the second question or follow-up question in a pair asks students to select from a set of multiple choice options their reasoning for their first answer. To redesign this question set to improve reliability, we essentially created more follow-up questions. When students answered correctly, they were given a "challenge" question, which targeted the same concept in a more difficult context. Also, when students responded to one of the original questions with unknown reasoning, they were given an analogous question, instead of repeating the same question with feedback. This had the benefit of removing cycles[ii], improving measurement concerns of prior exposure to the same item in the set.

There were three important concerns with this redesign. First, a reasonable student path through the set could now take up to twice as long. This time issue could be changed by starting high performing students farther into the bundles of questions, but this was not done in this case, in part given the experimental nature of the situation. Also higher performing students, who would likely be given the most additional question parts in the redesigned set as compared to the prior set, often more quickly complete work than other students, which might reduce the time impact of extending the sets even without

adjusting the set entry points by performance ability. However, to alert teachers to this concern and ensure that they allocate sufficient classroom time, the set was marked in DIAGNOSER as "experimental." Secondly, the final feedback score to students is usually presented as the number of questions they scored correctly out of the number attempted. Clearly, average scores in the face of the challenge questions would be lower, and possibly distressing for students who are anxious about their performance. Teachers were warned of this possibility through email as well. Finally, the current DIAGNOSER system does not allow questions to diagnose facets within more than one cluster. Some of the "challenge" questions we thought of and dismissed involve integrating ideas that are currently in different facet clusters. This is a design decision within DIAGNOSER, but one that might need to be revisited if the facet approach is to scale to more complex and interrelated content.

Figure 5 shows the Wright Map for the new item set, given to 968 students. The population of students studied was similar to the original context. Overall instrument reliability for the set was improved considerably, to .64 from .47. While this still falls short of a .8 target, and more revisions are shown to correct this in the following sections of this paper, the improvement from just the redesign of the question set is substantial. Figure 5 shows that there are nearly twice as many items and facets measuring in the 40s/50s and above as compared to the original set. While there still remain many facets measuring below the majority of the distribution, the locations of the persons and items are more aligned, changing the standard errors as shown in Figure 6 and improving the overall reliability. The standard error plot shows that in the redesigned set, only about 25 percent of students fall into high standard error regions as compared to 75 percent in the

original set. Standard errors begin to rise substantially beyond the rest of the distribution for student performance above about .6 logits, a part of the distribution that includes about 238 of 968 students, or about 25 percent of students. This compares to 75 percent of students falling into the higher standard error regions for the original Diagnoser question set.

The redesign of the item set and the elimination of cycles in favor of checking answer responses with a different second followup probe allowed us to make an interesting observation. Facet sequences that showed a propensity to go from answering a question correctly but then incorrectly answering the challenge item, or extension question, tended to have an overall threshold difficulty on the Wright Map that fell into the transition of facets from 70s to 50s. In comparison, facet sequences that did not show the behavior of flip-flopping from correct answers to wrong answers on extension questions tended to have higher threshold difficulty estimates, in the range of the 50s/40s facets or higher. An explanation for this is that when students show facets that move between right and wrong answers, they are at a lower proficiency on the construct than students who consistently answer in correct facets, thus the mixed facet sequences calibrated at a lower difficulty estimate.

The Wright Map in Figure 5 also shows that more item/facet pairs showed up as somewhat out of level based on the predictions of their facet locations. Note that the previous outliers of 7(41) and 4(51) remained out of level in this new analysis, but a few of the redesigned items showed new outliers at the higher levels, as shown in the outlier column of the Figure 5 Wright Map. This is not unusual when new items are tried out or item redesigns occur, such as the introduction of new challenge items that have not been

tested previously. Items can be reviewed for why the new outliers are happening and revised as deemed appropriate by subject matter experts, or assumptions regarding the theoretical aspects of the measurement goals or theory can be revised.

An additional strategy to improve instrument reliability

To further improve reliability, one approach might be to continue to extend items at the higher end or in other ways to more fully align the distribution of facets and persons. This might continue to lower the standard errors and increase the overall reliability of the question set. In this section, we briefly test a simulation of such data. The basic approach is to use the parameters of the estimated model to simulate two new sets of data. Both are based on the original item and step difficulty pattern. The first data set simulates data in which the student performance estimates are not fully aligned with the difficulty range of the score levels. The alignment selected for this first simulation is similar to the alignment of persons and items in the previous section, and reliability is thus expected to be in a similar range of about .6. The second data set simulates a more "ideal" alignment between persons and the range in which items measure, to see if reliability improves. We can use this second simulation to get an estimate of how much we might be able to improve reliability if the targeting of the item levels were better. This would help support the view, based just on simulated data, that further alignment might improve reliability in actual cognitive diagnoser function.

So first we simulated data, using ACER Conquest (Wu, Adams, & Wilson, 1998) to produce the less well aligned simulated data of students performing in a range similar to the analysis in previous example, for which some alignment but not full alignment is seen. EAP/PV reliability did not improve with the modeled data but dropped somewhat to

about .56, so remained in the .6 range as expected. However, in the second run that simulated data more ideal for the range of item and step difficulties, reliability improved to .83. Figure 7 shows a standard error plot for the second simulation.

While this result is good, and points to better alignment of persons and items as a solution for further improving reliability, simulated data does not necessarily reflect the performance of actual data, so it would be important to test this approach further with actual student data.

Conclusion

Cognitive diagnosers and other e-learning products that include embedded assessments are rapidly emerging. The intent is to collect assessment data to build a picture of some aspect of a learner's thinking. One concern for this area of e-learning and assessment is whether the diagnostic conclusions of such products are based on sound measurement evidence, including whether or not the diagnostics are reliable (Pellegrino, 2005; Pellegrino, Baxter, & Glaser, 1999; Pellegrino, Chudowsky, & Glaser, 2001). A reliability index helps quantify the impact that measurement error at the individual level may have on the accuracy of the inference (Adams, 2006). Low test reliability can suggest that less confidence can be placed in the diagnosis. This paper investigates some reliability evidence for one cognitive diagnostic product and offers some simple solutions for improving reliability with only minimal increases in the number of tasks and questions completed. Solutions shown include providing questions of appropriate difficulty to maximize item information across the distribution and including both pretest and post-test data during instrument calibration.

This paper also shows that the achievement of a particular reasoning facet on a single assessment question or task is unlikely to yield a reliable estimate of an expected reasoning facet on any other one item, which is the case for almost all assessment instruments. Assessments tend to require considerably more than one piece of student data to make a reliable estimate. However, a small group of items well aligned to student performance look promising for reliable estimation.

Note however that the question still remains of when reliability regarding assessment instruments is important in e-learning products. Whether used summatively or formatively, if a score report is made or information is used for course placement, learning intervention or adaptive opportunities for learning, it seems appropriate that measures should be reliable reflections of a student's performance. As mentioned previously for dynamic feedback, if the feedback is made based only on the last student response it seems reasonable that a reliable aggregated score may not be necessary. It is usually instructionally appropriate to give students feedback on their responses to questions, whether or not another response at another time or in another situation might have been different.

Sometimes the argument is made that reasonable reliability is necessary only for summative assessments. However, the acceptance of low reliability in e-learning instruments can be particularly troubling if systems award students with credit for "achieving" proficiencies in certain areas based on an observation in one session and then promptly take the credit away from students in the next session, even if this information is used formatively. Students can be left anxious and concerned about what this means about their performance, for instance, how they could have "forgotten" what they "knew."

But, in fact, if the instruments have low reliability, the mastery inference may have been questionable in the first place. For students, lack of control in being able to effectively monitor their own learning, especially if supposedly sophisticated measurement systems are providing unreliable results that students and teachers use to monitor performance, might have serious learning, metacognition and engagement consequences. These might include moving some students from mastery learning patterns to helpless learning patterns (Dweck & Leggett, 1988). Questions of fairness, metacognition, and effective teaching and learning strategies argue against endorsing instruments with known low reliability. Thus investigation of simple and effective solutions to improve the quality of assessment evidence in cognitive diagnosers and other promising e-learning products may make these products even more helpful to the instructor and to the learner.

Biographical Notes

Dr. Tara Madhyastha's interests include psychological modeling and educational data mining. She was previously a computer science professor at the University of California, Santa Cruz before she began working at Facet Innovations. She currently holds an affiliate appointment in the department of psychology at the University of Washington.

Dr. Jim Minstrell is a Senior Research Scientist and co-founder of FACET Innovations. He spent 30 years teaching mathematics, physics and integrated science and mathematics at the high school level. Part way into his teaching career Jim became aware of "misconceptions" research. He began studying the learning and teaching of his own students and those of his teacher colleagues. The negative connotation of the term misconceptions and the realization that there were strengths as well as problematic thinking in his students' understandings prompted Jim and colleagues to coin the term "facets of student thinking." Since retiring from the classroom, Jim has remained very active in science education. He has been a PI or co-PI of various research and development projects related to assessment, professional development and the Diagnoser project.

Dr. Kathleen Scalise is an assistant professor at the University of Oregon, and works in the areas of STEM e-learning and assessment. She is also co-director for the UC Berkeley Formative Assessment Delivery System project. She specializes in analysis of

student learning trajectories and learning outcomes in e-learning products, computer-adaptive approaches for assessment, and dynamic delivery of differentiated content. She served with the Curriculum Frameworks division of the California Department of Education for the California Science Framework for K-12 Public Schools.

Dr. Mark Wilson is a professor in the Graduate School of Education at the University of California, Berkeley. Wilson's work spans a range of issues in measurement and assessment from the development of new statistical models for analyzing measurement data, to the development of new assessments, to policy issues in the use of assessment data in accountability. Wilson recently chaired a National Research Council committee on assessment of science achievement, and is founding editor of a new journal, Measurement: Interdisciplinary Research and Perspectives.

References

Adams, R. (2006, April 5-7). *Reliability and Item Response Modelling: Myths, Observations and Applications.* Paper presented at the 13th International Objective Measurement Workshop, Berkeley, CA.

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

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

Dimitrova, V. (2002). *Interactive cognitive modelling - potential and challenges.* Paper presented at the Workshop on Individual and Group Modelling Methods that Help Users Understand Themselves, held in conjunction with ITS2002, San Sebastian, Spain.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256-273.

Gifford, B. R. (1999). *Computer-mediated instruction and learning in higher education: Bridging the gap between promise and practice.* Paper presented at the 104th Annual Meeting of the North Central Association of Colleges and Schools, Commission on Institutions of Higher Education (http://ishi.lib.berkeley.edu/sche/projects/university/april12gifford.html), Chicago, IL.

Gifford, B. R. (2001). Transformational Instructional Materials, Settings and Economics. In *The Case for the Distributed Learning Workshop*. Minneapolis, MN: The Distributed Learning Workshop.

Halloun, I. A., & Hestenes, D. (1985). Common-sense concepts about motion. *Am. J. Phys., 53*, 1056-1065.

Hopkins, D. (2004). *Assessment for personalised learning: The quiet revolution.* Paper presented at the Perspectives on Pupil Assessment, New Relationships: Teaching, Learning and Accountability, General Teaching Council Conference, London, England.

Hunt, E., & Minstrell, J. (2004). *The DIAGNOSER Project: Formative Assessment in the Service of Learning.* Paper presented at the International Association for Educational Assessment, Philadelphia.

Kennedy, C. A., Bernbaum, D. J., Timms, M. J., Harrell, S. V., Burmester, K., Scalise, K., & Wilson, M. (2007). *A Framework for Designing and Evaluating Interactive E-Learning Products.* Paper presented at the 2007 AERA Annual Meeting: The World of Educational Quality, Chicago.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn & T. Okada (Eds.), *Designing for Science: Implications for Professional, Instructional, and Everyday Science*. Mahwah: Lawrence Erlbaum Associates.

Minstrell, J., Anderson, R., Minstrell, J., & Kraus, P. (in preparation). Bridging from Practice to Research and Back. In W. Binder & C. Sterns (Eds.), *Assessment*: NSTA Press.

National Research Council/National Academy of Sciences. (1996). *National Science Education Standards*. Washington, D.C.: National Academy Press.

Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats*. Norwell, MA: Kluwer Academic Publisher.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative Item Types for Computerized Testing. In W. Van der Linden, Glas, C. A. W. (Ed.), *Computerized Adaptive Testing: Theory and Practice* (pp. 129-148). Norwell, MA: Kluwer Academic Publisher.

Pellegrino, J. W. (2005). Discussant for Moving Technology Up-Design Requirements for Valid, Effective Classroom and Large Scale Assessment, *2005 Annual Meeting American Educational Research Association (AERA), Demography & Democracy in the Era of Accountability*. Montreal, Canada.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines" Problem: Linking Theories of Cognition and Learning with Assessment and Instructional Practice. *Review of Research in Education, 24*, 307-353.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. In N. R. C. Center for Education (Ed.). Washington, D.C.: National Academy Press.

Rosenbaum, P. R. (1988). Item Bundles. *Psychometrika, 53*, 349-359.

Scalise, K., Bernbaum, D. J., Timms, M. J., Harrell, S. V., Burmester, K., Kennedy, C. A., & Wilson, M. (2006). Assessment for e-Learning: Case studies of an emerging field [Electronic Version]. *13th International Objective Measurement Workshop*. Retrieved April 5 from http://bearcenter.berkeley.edu/publications/Scalise_eLearning.pdf.

Scalise, K., & Claesgens, J. (2005). *Personalization and customization in new learning technologies: Getting the right assets to the right people.* Paper presented at the Demography and Democracy in the Era of Accountability, American Educational Research Association Conference, Montreal, Canada.

Tatsuoka, K. K., Birenbaum, M., & Arnold, J. (1989). On the Stability of Students' Rules of Operation for Solving Arithmetic Problems. *Journal of Educational Measurement 26*(4), 351-361.

Taylor, C. R. (2002). E-Learning: The Second Wave. Retrieved July 10, 2006, from http://www.learningcircuits.org/2002/oct2002/taylor.html

Tomlinson, C. A., & McTighe, J. (2006). *Integrating Differentiated Instruction + Understanding by Design: Connecting Content and Kids*. Alexandria, VA: Association for Supervision and Curriculum Development.

Trivantis. (2005). Present Day Custom eLearning. Retrieved July 12, 2006, from http://www.trivantis.com/custom-elearning/custom-elearning.htm

Turker, A., Görgün, I., & Conlan, O. (2006). The Challenge of Content Creation to Facilitate Personalized E-Learning Experiences. *International Journal on E-Learning, 5*(1), 11-17.

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

Wilson, M., & Adams, R. J. (1995). Rasch Models for Item Bundles. *Psychometrika, 60*(2), 181-198.

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* (pp. 89-108). Arlington, VA: NSTApress.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M., Adams, R. J., & Wilson, M. (1998). The generalised Rasch model. In *ACER ConQuest*. Hawthorn, Australia: ACER.

Figure 1. Flow chart for part of a Diagnoser question set

Figure 2. Objectives for Describing Motion and Determining Speed, Alignment with the National Standards (NSES) and Benchmarks (BSL)

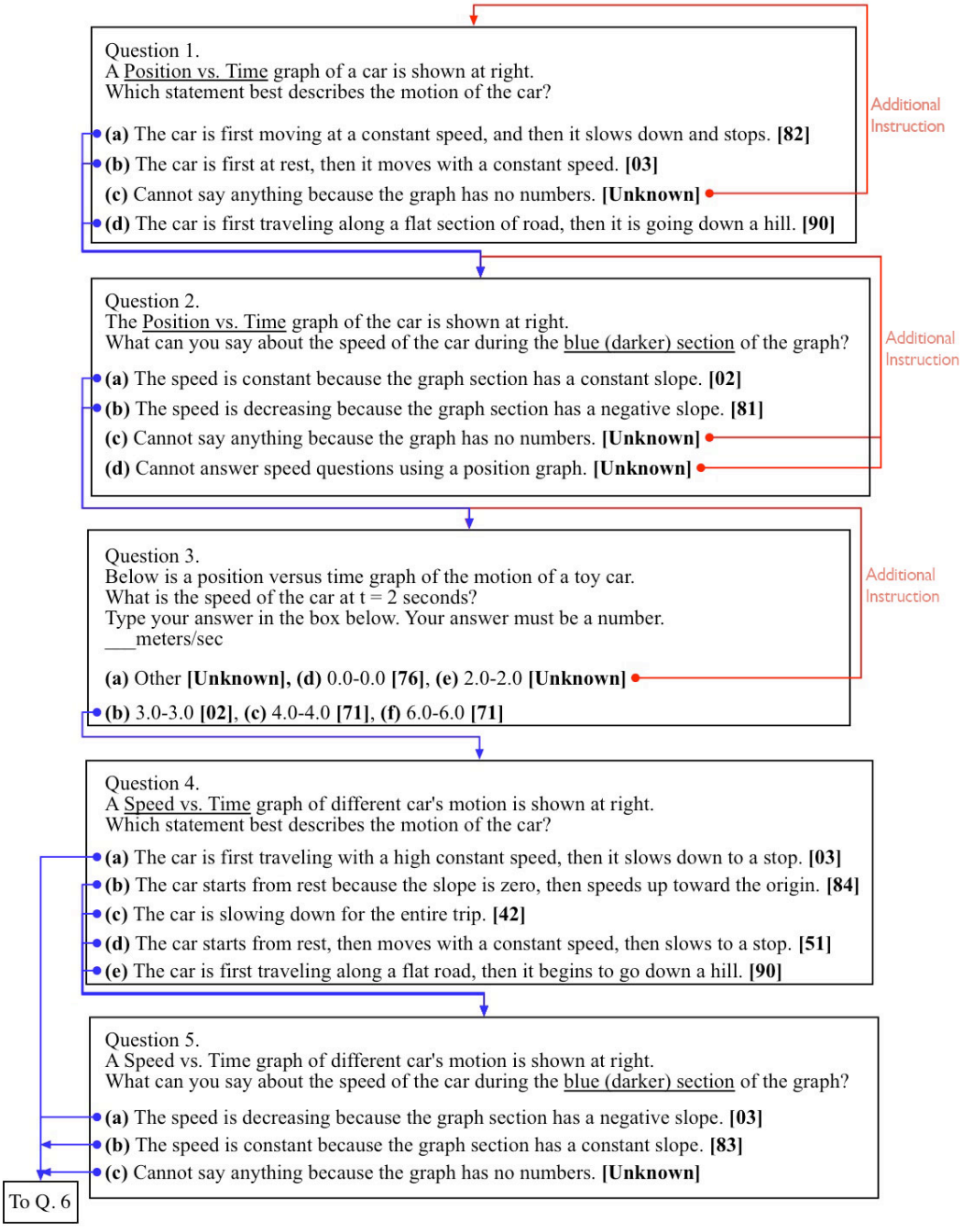
Figure 3. A Wright Map for the Diagnoser example question set and student sample

Figure 4. Standard error plot for the original DIAGNOSER example question set and student sample, with student performance estimates showing in the logit scale of the Figure 3 Wright Map.

Figure 5. A Wright Map for the revised DIAGNOSER example question set

Figure 6. Standard error plot for the revised Diagnoser question set, with student performance estimates in the logit scale of the Figure 5 Wright Map.

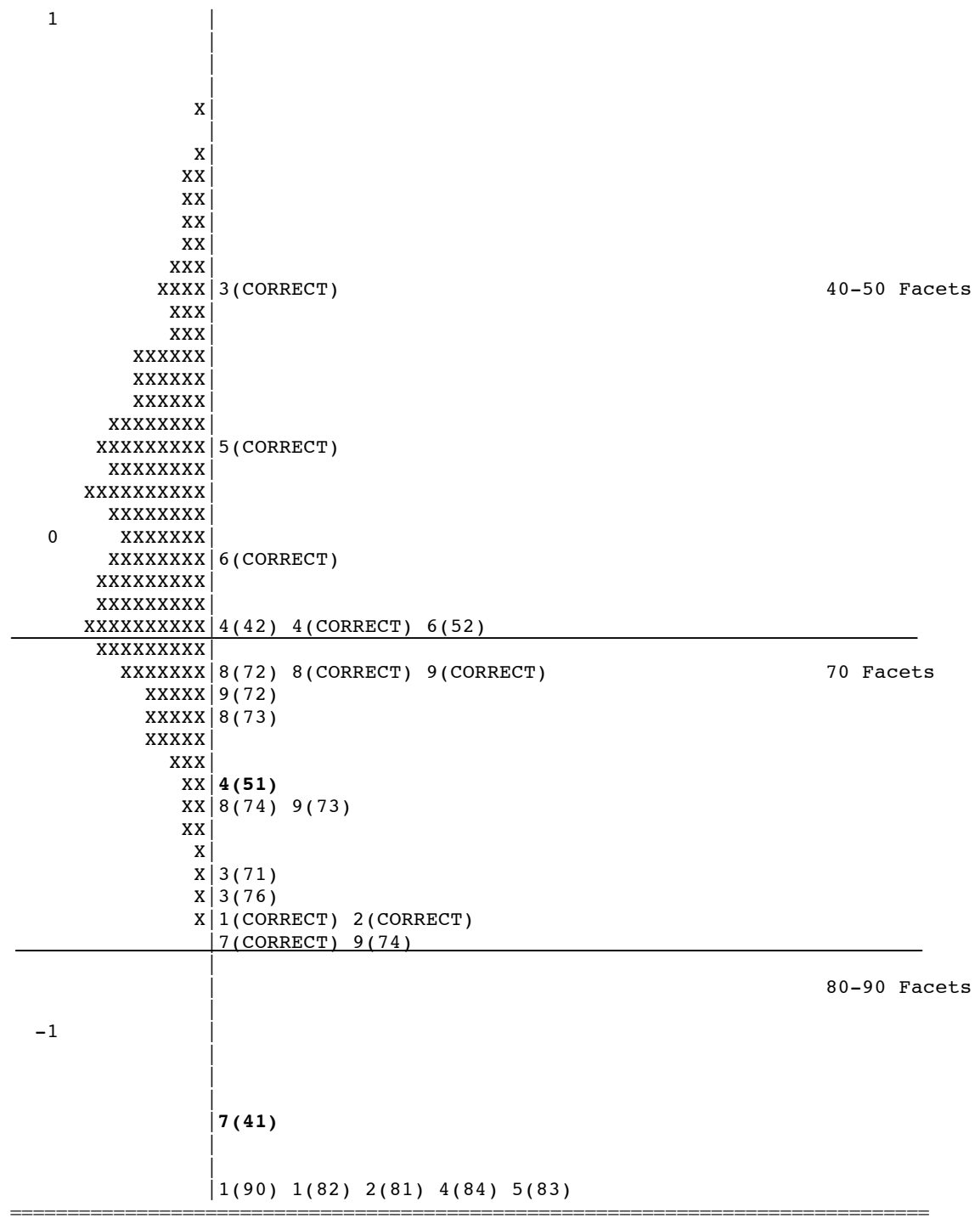
Figure 7. Standard error plot for the revised Diagnoser question set with simulated pretest data included in the calibration.

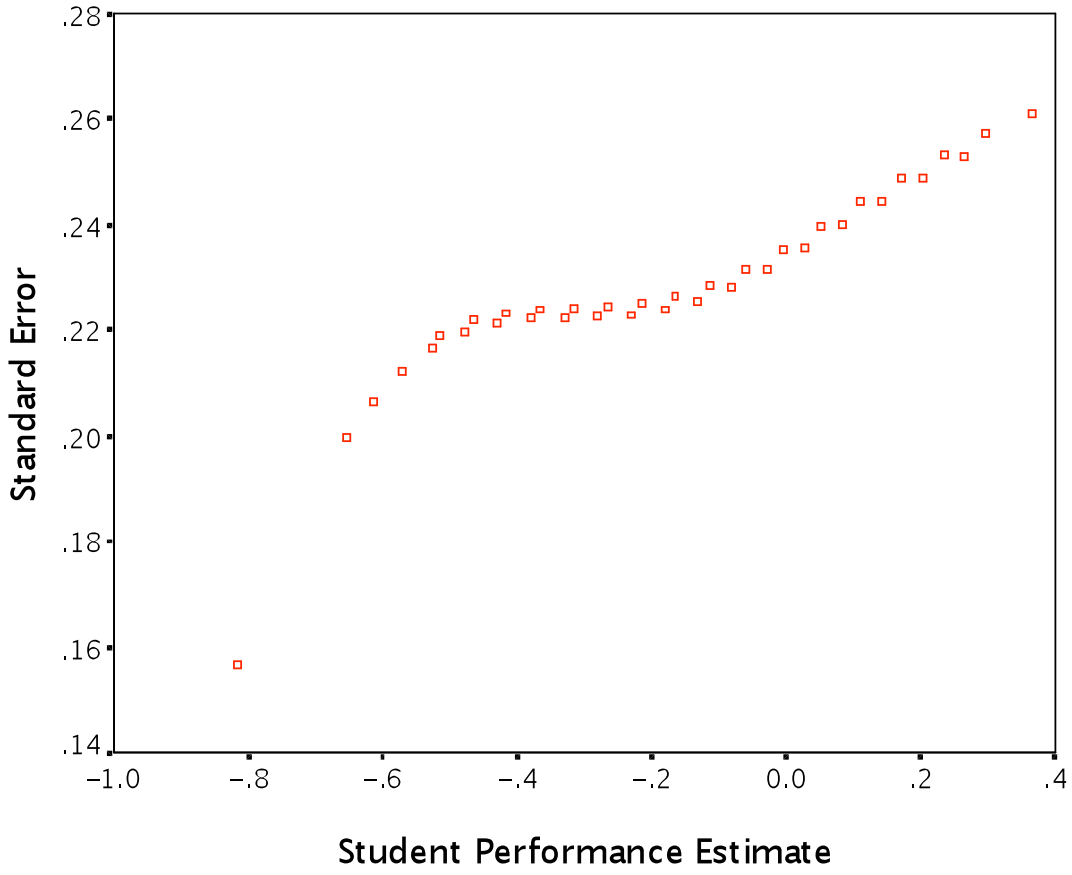

Question 1.
A Position vs. Time graph of a car is shown at right.
Which statement best describes the motion of the car?
(a) The car is first moving at a constant speed, and then it slows down and stops. [82]
(b) The car is first at rest, then it moves with a constant speed. [03]
(c) Cannot say anything because the graph has no numbers. [Unknown]
(d) The car is first traveling along a flat section of road, then it is going down a hill. [90]
Additional Instruction
Question 2.
The Position vs. Time graph of the car is shown at right.
What can you say about the speed of the car during the blue (darker) section of the graph?
(a) The speed is constant because the graph section has a constant slope. [02]
(b) The speed is decreasing because the graph section has a negative slope. [81]
(c) Cannot say anything because the graph has no numbers. [Unknown]
(d) Cannot answer speed questions using a position graph. [Unknown]
Additional Instruction
Question 3.
Below is a position versus time graph of the motion of a toy car.
What is the speed of the car at t = 2 seconds?
Type your answer in the box below. Your answer must be a number.
___meters/sec
(a) Other [Unknown], (d) 0.0-0.0 [76], (e) 2.0-2.0 [Unknown]
(b) 3.0-3.0 [02], (c) 4.0-4.0 [71], (f) 6.0-6.0 [71]
Additional Instruction
Question 4.
A Speed vs. Time graph of different car's motion is shown at right.
Which statement best describes the motion of the car?
(a) The car is first traveling with a high constant speed, then it slows down to a stop. [03]
(b) The car starts from rest because the slope is zero, then speeds up toward the origin. [84]
(c) The car is slowing down for the entire trip. [42]
(d) The car starts from rest, then moves with a constant speed, then slows to a stop. [51]
(e) The car is first traveling along a flat road, then it begins to go down a hill. [90]
Question 5.
A Speed vs. Time graph of different car's motion is shown at right.
What can you say about the speed of the car during the blue (darker) section of the graph?
(a) The speed is decreasing because the graph section has a negative slope. [03]
(b) The speed is constant because the graph section has a constant slope. [83]
(c) Cannot say anything because the graph has no numbers. [Unknown]
To Q. 6

Speed Set Objectives

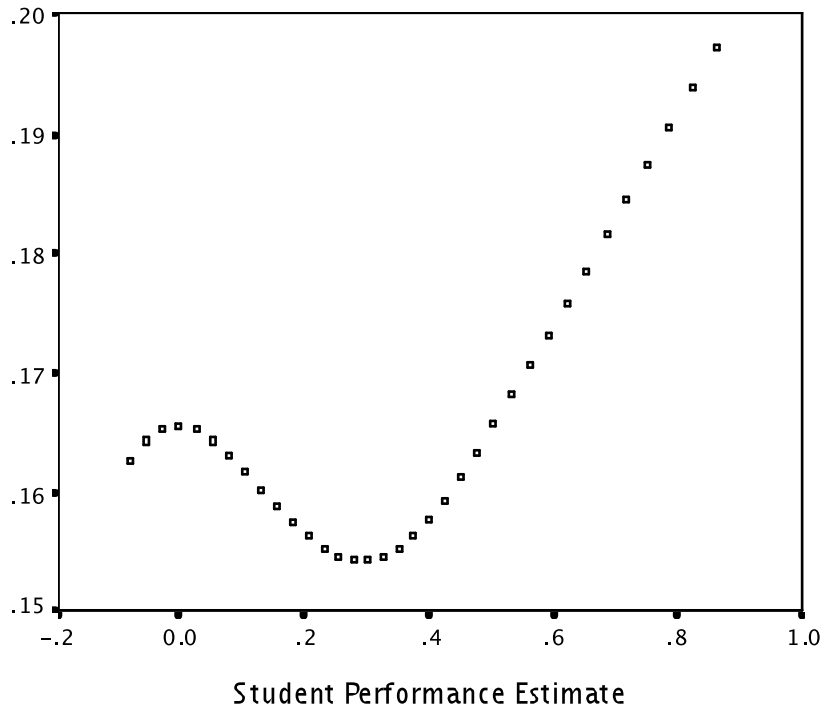
- The motion of an object can be described by its position, direction of motion, and speed. That motion can be measured and represented on a graph. (NSES p154, grades 5-8)
- Students should continue describing motion. And they can be more experimental and quantitative as their measurement skills sharpen. Determining the speed of fast things and slow things can present a challenge that students will readily respond to. (BSL 4F, grades 3-5 Text discussion of student learning)
- Graphs can show a variety of possible relationships between two variables. (BSL 9B, grades 6-8)
- Mathematics is the study of any patterns or relationships, whereas natural science is concerned only with those patterns that are relevant to the observable world. (BSL 2A, grades 9-12)
- Tables, graphs, and symbols are alternative ways of representing data and relationships that can be translated from one to another. (BSL 9B, grades 9-12)

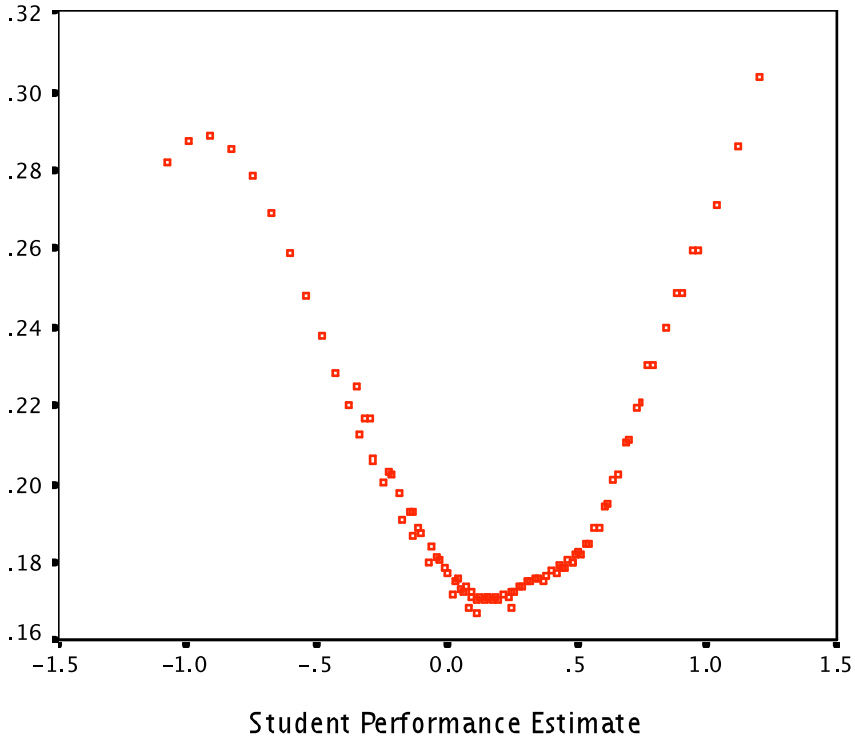
MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS




.28
.26
.24
.22
.20
.18
.16
.14
Standard Error
−1.0
−.8
−.6
−.4
−.2
0.0
.2
.4
Student Performance Estimate

```
Minstrell May 3 2006 Run 1, PC     Tue Mar 07 20:25:45              Outlier              ~Levels
MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS
=========================================================
     Generalised-Item Thresholds (0 is Mean of Items)
---------------------------------------------------------
            X|
             |
            X|
1           XX|
            XX|
             X|
            XX|
           XXX|
          XXXX|
           XXX|2(Correct/Correct)                                  2(Correct/70)
           XXX|
         XXXXX|
          XXXX|
         XXXXX|3(Correct 0r -1/Correct)
        XXXXXX|
       XXXXXXX|5(Correct/Correct)                                                 Tendency all correct
       XXXXXXX|
       XXXXXXX|5(Correct/40) 9(Correct)
       XXXXXXX|1(Correct/Correct) 6(Correct) 9(41)                               Threshold 50s to 40s
-------------------------------------------------------------------------------------------------------
         XXXXX|1(Correct/50)
       XXXXXXX|1(Correct/90) 1(Correct/80) 2(Correct/80)
       XXXXXXX|1(-1/Correct) 4(Correct/76) 4(Correct/Correct) 6(52)   1(-1/80)
      XXXXXXXX|2(-1/52) 2(-1/Correct) 5(Correct/84)                              Threshold 70s to 50s And
    XXXXXXXXXX|4(Correct/-1) 4(Correct/80)                                      Threshold correct/wrong
-------------------------------------------------------------------------------------------------------
       XXXXXXX|                                                    4(42)
       XXXXXXX|8(72) 8(Correct)
           XXX|
          XXXX|9(70)
          XXXX|8(73)
           XXX|
0           XX|
            XX|
             X|
             X|
             X|3(76 Or -1/76) 3(71 Or -1/71) 5(83) 8(74) 9(73)              Threshold 80s to 70s
-------------------------------------------------------------------------------------------------------
             X|9(80)                                                 4(51 Or CORRECT/51)
              |9(83)
              |
              |
              |
              |
              |
              |7(Correct)
              |
              |
              |
              |9(84)
              |
              |1(82 Or -1/82) 2(81 Or -1/81) 4(84) 9(90)             7(41)   Threshold 90s to 80s
-------------------------------------------------------------------------------------------------------
Each 'X' represents 6.8 cases
```


.20
.19
.18
.17
.16
.15
-.2
0.0
.2
.4
.6
.8
1.0
Student Performance Estimate



[i] The EAP/PV test reliability indice is an internal consistency reliability estimate similar to Cronbach's alpha. It is based on the item response model estimates rather than the raw score, and it is useful in situations where there is too much planned missing data for Cronbach's alpha to be estimated. Computer adaptive diagnostics "plan in" missing data, since students skip or branch through questions depending on their answers, so this is a common occurrence in e-learning.

[ii] "Cycles" occur when students answer a question incorrectly, are given a learning exercise or some other feedback, and then are asked the same question again. This is called a cycle because students "cycle back" to the same question.