Tech., Inst., Cognition and Learning, Vol. 3, pp. 151-166
Reprints available directly from the publisher
Photocopying permitted by license only

Diagnostic Assessment in Mathematics Problem Solving

VALERIE SHUTE* & JODY UNDERWOOD

Educational Testing Service, Princeton, NJ

The United States has recently seen falling test scores for mathematics problem solving in comparison to other countries. This paper reviews current approaches to diagnosing mathematics problem solving, and then introduces emerging technologies being developed at Educational Testing Service that address open areas found in the review. The application of these technologies to assessment design must be weighed against concerns for construct validity, equity, and access. The validity question is addressed by using evidence-centered design (ECD) methods to build an evidentiary argument. An innovative project called Mathematics Intervention Module (MIM) for helping students improve their mathematical problem solving skills is described that uses ECD methods in concert with the emerging technologies, with a focus on diagnosis, feedback, practice, and items that integrate targeted knowledge and skills.

Keywords: Assessment, cognitive diagnosis, evidence, instructional intervention, mathematics problem solving, proficiency model, tasks.

If you only have a hammer, you tend to see every problem as a nail. —Abraham Maslow

THE GENERAL PROBLEM

America's 15-year-olds performed below the international average in mathematics literacy and problem solving, according to the latest results

*Corresponding author: Vshute@Ets.Org

from the Program for International Student Assessment (PISA). The test, given in the spring of 2003, assesses the ability of 15-year-old students from various countries (including 30 of the most developed) to apply learning to problems with a real-world context (see PISA Report, 2004). Students in the following countries outperformed the United States in mathematics literacy in 2003: Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Hong Kong-China, Iceland, Ireland, Japan, Korea, Liechtenstein, Luxembourg, Macao-China, Netherlands, New Zealand, Norway, the Slovak Republic, Sweden, and Switzerland.

These same 23 countries, plus Hungary and Poland, outperformed the United States in mathematics problem solving. *U.S. 15-year-olds scored measurably better than their counterparts*[1] *in only 3 of 30 nations on the new international test of problem solving in math.* Moreover, the U.S. has the poorest outcomes per dollar spent on education. In short, U.S. students are performing poorly on mathematics tasks that involve transfer of learning and problem solving skills. We need to bolster our students' problem solving skills to compete effectively internationally, in the near future.

The purpose of this paper is to examine ways to improve math problem solving, focusing our attention on the role of careful *diagnosis*. We begin by reviewing some systems that assess math problem solving skills.

CURRENT DIAGNOSTIC APPROACHES

What features of the student, task, content, or environment are important to analyze in order to diagnose strengths and weaknesses and thus support learning? How are current diagnostic and tutoring systems addressing these issues? Here are a few well-known systems that diagnose students' mathematics problem solving skills and help students acquire relevant content. The systems differ in terms of measurement technique employed (e.g., percent correct scores on 1-3 tasks vs. Bayesian inference networks), but perhaps more importantly, on the level of the diagnosis—*local* (e.g., analyzing specific problem-solving steps) or *global* (e.g., inferring mastery status of general proficiencies).

[1] "Counterpart" refers to similar students, based on age and grade, who reside in different countries. For more on international standing and comparison among NAEP, TIMSS, and PISA results, see http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf.

- *ALEKS* is a web-based system that applies knowledge space theory developed for K-12 math (Falmagne, Cosyn, Doignon, & Thiery, 2004). It uses a model of precedence for mathematics concepts (problem types), and through assessment, determines what a student knows and what he or she is ready to learn (global diagnosis). The strength of ALEKS is in its broad diagnostic ability. Local level diagnostics are not quite as complete. For instance, its sensitivity to student input at the question level is limited, although an important part of human tutoring occurs at this level.
- Carnegie Learning's *Cognitive Tutors* provide interactive environments via problem solving and worksheet-like activities (Corbett, McLaughlin, & Scarpinatto, 2000). For example, in algebra, students not only solve structured problems using equations, they also fill in tables relating variables to each other, providing a concrete basis on which to construct equations. In all interactions, a cognitive model follows the students' problem solving efforts, providing immediate feedback if the student strays. Students receive two types of feedback: help to keep them on track as they solve a problem (strong focus on local level diagnosis), and a summary of estimated mastery on relevant skills (global level diagnosis). What's missing, however, is a model of how component skills are linked to each other and a clear indication of what a student is ready to learn. Instead, the tutor follows a fairly fixed curriculum upon which the problems and worksheets are built and presented.
- *Ms. Lindquist* is designed to teach students to write algebra expressions for word problems (Heffernan & Koedinger, 2002). It asks students to solve open-ended problems where the answer is an expression that involves numbers, variables, and the four basic arithmetic operations. Student modeling and problem selection strategies are rather simple—the student has to answer three problems in a row correctly to move to the next lesson. The problems seem to be represented internally as subparts with English prose associated with each sub-part. The system tries to recognize correct parts of the answer in order to employ different tutoring strategies (local level diagnosis). There does not seem to be a representation of knowledge and skills that link the problems to each other given the linear presentation of problems.
- *ActiveMath* is a web-based learning system that dynamically generates interactive math "courses" adapted to the student's goals, preferences, capabilities, and knowledge (Melis & Andrés, 2002). ActiveMath keeps track of student progress and offers advice while students interact with the

> system. The content is represented in a format that includes such elements as types of exercises and difficulty of problems. Student model information, navigation information, reading time, and assessed performance are used by *Suggestors* to diagnose possible difficulties and provide appropriate feedback (local level diagnosis). Links between underlying concepts are not apparent, so there is no way to select appropriate next problems for a student to solve.

As shown above, some systems are diagnostically effective at the local level while others focus more on global diagnosis. Only a few systems provide the basis for a range of diagnostics. For instance, AuthorIT (Scandura, 2005) diagnoses local (e.g., "atomistic" skills) and global (e.g., problem solving skills) knowledge based on abstract syntax tree (AST) representations of lower and higher order knowledge, respectively. This paper will present another possible solution that attends to both levels, via application of an evidence-centered design (ECD) approach. It accomplishes this by providing a bridge from observables (performance data) to unobservables (proficiencies) via evidence models—at multiple levels.

TECHNOLOGY SOLUTIONS

To diagnose student input at local levels and infer proficiency status at global levels, we are employing a variety of technological solutions in our assessments, such as: automated scoring of different constructed response types, item generation, adaptive testing, and the capability to present or simulate "authentic" problem solving contexts. Each of these must be weighed against concerns for construct validity, equity, and access (Bennett & Bejar, 1998; Shute, Graf, & Hansen, 2005).

New directions in educational and psychological measurement are also allowing more accurate estimations of students' proficiencies. Consequently (albeit slowly), assessments are evolving. Advanced technologies are permitting us to administer embedded (stealth) assessments during the learning process, extract ongoing, multi-faceted information from a learner, and react in immediate and helpful ways. These new assessments can be accomplished via automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans (e.g., estimating proficiencies across a network of skills). The big question is not about collecting this rich digital data stream, but rather, how to make sense of what can potentially become a deluge of information.

EVIDENCE-CENTERED DESIGN

Evidence-centered design (ECD; e.g., Mislevy, Steinberg, & Almond, 2000; 2003) is a viable solution to this quandary. It provides (a) a way of reasoning about assessment design, (b) a way of reasoning about examinee performance, (c) a data framework of reusable assessment components, and (d) the means to unify and extend probability-based reasoning to assessment (whether traditional standardized tests, classroom tests/quizzes, simulations, gaming environments, portfolios, etc.).

How does ECD work? The key idea of ECD is to specify the structures and supporting rationales for the evidentiary argument of an assessment. By making the evidentiary argument more explicit, the argument becomes easier to examine, share, and refine. Argument structures encompass, among other things, the claims (inferences) one wishes to make about a student, the observables (performance data) that provide support for those claims, the task performance situations that elicit the observables from the students, and rationales for linking it all together. The three main models used in ECD are:

- *Proficiency Model*: Establish claims about a particular piece of knowledge, skill, or ability. The proficiency model describes what is to be measured, conditions under which the ability is demonstrated, and the range and relations of proficiencies in the content area.
- *Evidence Model*: Define evidence needed to support claims. Evidence models describe what is to be scored, how to score it, and how to combine scores into claims. These models thus establish the boundaries of performance and identify observable actions that are within those boundaries.
- *Task Model:* Identify tasks that are able to elicit that evidence. Task models specify the inputs required to perform the observable actions as well as the work products that result from performing the observable actions.

DIAGNOSIS

To figure out the nature and extent of a problem in a student's problem solving efforts, we need to design tasks such that this information can be disentangled and interpreted in valid and reliable ways. A good diagnostic system should be able to accurately infer proficiency estimates (i.e., levels of mastery) for a student. This process begins with the design of a reasonable (i.e., accurate and informative) proficiency model, which provides the basis for both local and global level diagnoses to occur. Information from students' interactions with tasks or problems will be analyzed to inform relevant proficiencies. These task-level diagnoses provide local support (via scoring rules and feedback) while estimates of proficiency provide the basis for selecting the next task for the student (via selection rules or algorithms—beyond the scope of this paper).

Proficiency estimates can assume various forms, from percent correct data to probabilistic estimates of mastery of knowledge or skills via either a Bayesian network or regression equations (e.g., Mislevy, Almond, Yan, and Steinberg, 1999; Shute, 1995). Our approach to diagnostic assessment rests on the belief that students must actually demonstrate knowledge/skill capability within carefully crafted and contextualized tasks. Thus, a key component in an assessment design is the provision of a rich set of activities in which learners can practice, receive targeted feedback, and demonstrate their level of performance.

Learning outcomes (e.g., objectives, standards, and what needs to be reported) can serve as a starting point for developing a proficiency model, and there should be a rich pool of activities from which the system may draw at any time and which can provide instruction, assistance, and feedback to the learner in addition to "just a summary score." In all cases, interpretation of proficiency is a function of the goodness of the evidence collected. In a valid proficiency model, each piece of knowledge, skill, and ability will be linked to more than one task so that evidence of a student's performance can be accumulated in a number of different contexts. In a hierarchical proficiency model, evidence of one skill's mastery can also feed into a parent or child skill's mastery estimation. An example proficiency model is presented later in the context of our Example Project.

EVIDENCE

Individual responses to assessment tasks, as well as patterns of responses, serve as the primary basis for evidence of proficiencies—locally at the task level, and globally at the proficiency estimation level. Information may be culled directly from the students' behaviors and work products as they interact with and complete items within an assessment task (or task set). Based on exactly what the student produces (i.e., evidence) in response to a given math problem-solving task, inferences can be made about the source of the problem or strength of a set of skills. Obviously, open-ended tasks will invoke more varied evidence than multiple-choice responses. ETS has been developing tools to analyze various response types, discussed within the following section.

EXAMPLE PROJECT

The name of the example project described in this section is MIM, for Mathematics Intervention Module. MIM is an online application designed to help students become proficient in the state mathematics standards. The initial focus is on Algebra I, but it may be extended to other subjects in subsequent releases. The module is based on a proficiency model that describes the skills that must be mastered to be judged proficient in a standard. Each module presents students with open-ended questions dealing with the various skills identified in the proficiency model. These questions require the student to respond with (1) a number, (2) an expression or an equation, (3) a graph, or (4) text, all of which are automatically scored.

Diagnostic Feedback. All responses in the intervention module are automatically evaluated, with immediate feedback provided to the student. Feedback is directed at the error that the student has made, and is not simply, “Wrong. Please try again.” Similar to a human tutor, MIM attempts to give some indication of why the student’s answer was wrong. The student is given three attempts to answer each question correctly, with progressively more detailed feedback provided along the way. The correct answer, with an associated rationale, is presented if the student answers incorrectly three times. In addition, if the student is judged to be in need, the module presents a short (i.e., 2-4 minute) instructional video that covers the problematic skill. These “instructional objects” reinforce the learning that is taking place as the student works through the questions and reads the feedback.

Instructional Objects. A specific instructional object (IO) is presented when students require all the three levels of feedback. There are currently about 16 IOs produced for the current MIM prototype. Within an IO, the flow of instruction proceeds as follows: (a) introduce the topic using concrete and engaging context, (b) state a particular problem that needs solving, (c) provide relevant definitions, (d) illustrate the concept within different examples (both prototypical and counter-examples), (e) provide sufficient practice and interactivity, and (f) conclude with summary and reflection screens.

Practice Opportunities. The teacher has the option of assigning multiple-choice questions for additional practice on each skill. The teacher can (a) require these practice questions of all students who seem not to have mastered the skill, (b) make the practice questions optional, or (c) configure the module so that the practice questions are not delivered.

Integrating Knowledge and Skills. The final section of each intervention module is a set of integrated open-ended questions that deal with a common theme or contextual situation. These questions reflect the standard as a whole. Like the open-ended questions earlier in the module, these integrated questions involve responses that require the entry of a number, an expression or an equation, a graph, or text.

Information to the Teacher. After the student completes an intervention module, the teacher receives a summary report. In addition, the teacher can review the student's entire session, viewing the student's responses to each question. Classroom summaries are also possible, so that teachers can see, at a glance, how their students are progressing on the targeted standard.

Proficiency Model. As described earlier, a proficiency model generally describes the skills that must be mastered to be judged "proficient" in relation to a specific standard, and displays the relationships between these skills. The initial MIM prototype uses a proficiency model that analyzes the standard, "Translate word expressions to symbolic expressions or equations and then solve and/or graph" (see Figure 1). By working down the model, one can see how the component skills are isolated.

In this standard, "word expressions" means information contained in a story, a contextual description, or some other real-life situation. At a high level, this standard can be divided into three parts, each corresponding to a separate skill and each represented by a node (three white ovals) on the model. The first skill is to translate the information given in the story into an equation or graph or some other symbolic expression. The second skill is to solve the equation, and the third is to graph the equation and obtain useful information from the graph. For the purposes of this model, we are assuming that the equations and graphs are linear.

The first skill (translate context to equations and/or graphs) can be further divided into several sub-skills. To translate contextual information into an equation or graph, one must first identify the variables, and then identify the operations (addition, multiplication, and so on) that connect the variables, and finally put it all together correctly to form the relevant

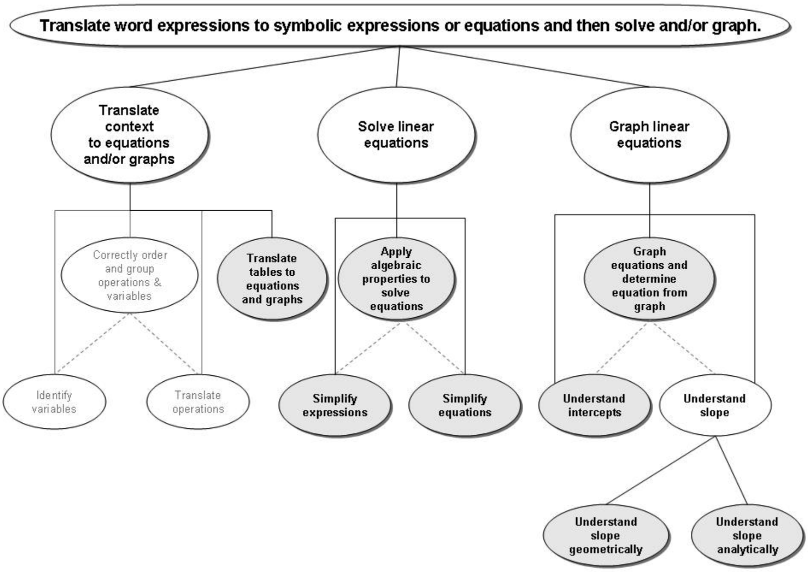

Translate word expressions to symbolic expressions or equations and then solve and/or graph.
Translate context to equations and/or graphs
Solve linear equations
Graph linear equations
Correctly order and group operations & variables
Translate tables to equations and graphs
Apply algebraic properties to solve equations
Graph equations and determine equation from graph
Identify variables
Translate operations
Simplify expressions
Simplify equations
Understand intercepts
Understand slope
Understand slope geometrically
Understand slope analytically

FIGURE 1
Simplified Proficiency Model for the MIM prototype.

equation. Each of these three skills is represented by a node within the model, and each node is connected to its parent node, *Translate context to equations and/or graphs*. In addition, dotted lines connect the third sub-skill with the first two because the third sub-skill involves the proper application of the first two.

In the current proficiency model, these nodes are faded. Due to constraints in the current project, we could not fully implement the mathematical content for these skills at this time. Instead, we teased out part of this content area and displayed it as a separate skill—entering contextual information into a table and then translating the table into a linear equation or graph. This skill is displayed as a gray node, indicating that this is one of the skills implemented in the current release of the intervention module.

A similar analysis applies to the second high-level skill (solve linear equations). This skill can be divided into three sub-skills: (1) use the rules of algebra to simplify expressions, (2) use the rules of algebra to simplify equations, and (3) combine the first two skills to solve equations.

Again, each of the three skills is connected to the parent skill. In addition, the third skill (apply algebraic properties to solve equations) is connected by dotted lines to the first two skills as it represents a proper application of the first two. All three of these nodes are displayed as gray because all three are implemented in the intervention module.

The third high-level skill (graph linear equations) is subdivided into three component sub-skills: understand intercepts, understand slope, and use knowledge of intercepts and slope to graph equations and determine equations from graphs. In addition, the Understand slope skill is further divided into two parts: Understand slope geometrically and Understand slope analytically. The "leaf nodes" (i.e., nodes with no children) are displayed as gray and are implemented in the intervention module.

The various elements of an intervention module—the open-ended questions, the instructional videos, and the multiple-choice practice questions—are presented to the student according to a carefully planned instructional design, based on principles of assessment and instruction that have been developed by researchers at ETS (Kuntz, et al., 2005). We used ECD (Evidence Centered Design) to develop the underlying proficiency model, scoring rules, and informative assessment tasks, and incorporated research-based features into MIM to support learning (e.g., timely diagnostic feedback, tailored content, and multiple representations of concepts).

In the following example, the integrated task set, as mentioned earlier, is presented at the end of the module, and its function is to assess the conjoined knowledge and skill elements. Finding a solution to the task requires the student to graph a line, find the equation of the line, identify the *y*-intercept and slope, state their significance in the context of the problem, and extrapolate data.

Music World Task. You found a new web site that claims to offer the best deal around for buying music CDs. The web site isn't clear about the cost for each CD or the cost of shipping and handling (except to say shipping is a flat fee), but it does give you the following information:

Number of CDs Ordered	**1**	**2**	**3**
Total Cost (with Shipping & Handling)	$9	$14	$19

1. Plot the data in the table on the graph (provided). Draw the line that contains the data points.

2. Assume that total cost is a linear function of number of CDs ordered.
   a. Write an equation of the line that contains the data points. Show your work.
   b. What is the slope of the line that contains the data points?
   c. What does that slope represent in the context of this problem?
   d. What is the *y*-intercept of the line that contains the data points?
   e. What does that *y*-intercept represent in the context of this problem?

3. Your friend says that he can get 15 CDs from the web site for $64.00. Is your friend correct? Explain.

In the *Music World* task, above, each node in the proficiency model may be linked, via different evidence models, to a number of tasks. As the student interacts with the system and answers questions, evidence is accumulated and the student model is updated. If a student demonstrates that she can calculate the slope using points on a graph, and interpret what it means in the context of the problem, the corresponding nodes in the proficiency model will show higher estimates of mastery. Moreover, because of the hierarchical nature of the proficiency model, the parent node, "Understands slope," may also automatically increase slightly. The converse is true for failing to solve the problem correctly. In general, proficiency information in the student model can highlight specific areas that need more instructional support.

To further facilitate the diagnosis of student performance, the system knows about a number of common misconceptions in relation to the skills in the proficiency model. To illustrate, in relation to the calculation and interpretation of the slope, some of the salient misconceptions and errors include inaccurate symbolic and graphical modeling of data, misunderstanding of slope as a rate of change, misinterpretation of slope and *y*-intercept in real contexts, and inability to use the equation of a line as a tool to predict linear behavior (i.e., extrapolation). These can be used as indicators to help diagnose the problems with the knowledge and skills in the proficiency model. A teacher or instructional module, armed with this information, can be considerably more effective in providing a targeted intervention.

Following are some general requirements for a student to get a maximum score per item element in the Music World example:

1. Graphs points correctly with respect to the axes.

2a. Writes a correct equation for the line based on an accurate reading of the graph or correct calculations using a linear form.

2b. Gives the correct slope based on the graph or the equation written in part 2a.
2c. *Gives a clear and correct interpretation of slope in context.*
2d. Gives the correct *y*-intercept based on the graph or the equation written in part 2a.
2e. Writes a clear and correct interpretation of *y*-intercept in context.
3. Writes an answer and justification that are correct, based on the equation given in question 2 or based on the graph in question 1.

Let's look at requirement 2c in more detail. The learning objective is that the student can give a clear and correct interpretation of slope in the context of the problem. The work product is a written (typed) response to an assessment item. The three levels are:

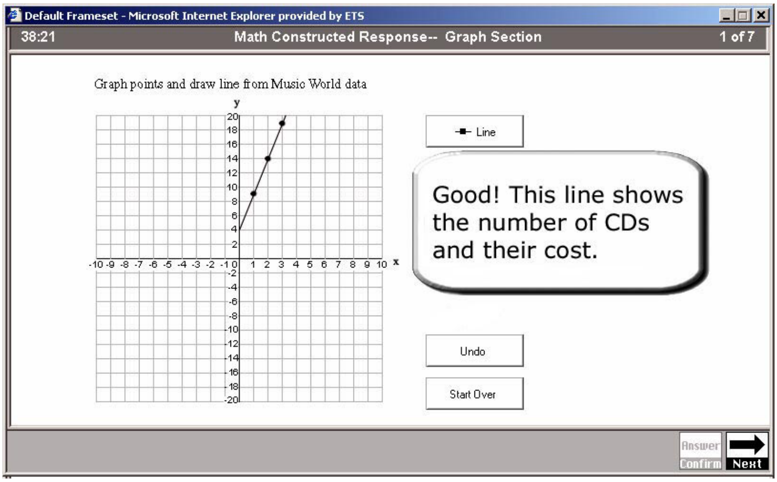
- *Low*: Student describes something that does not relate to the contextual variables related to slope (i.e., something other than CD price and shipping and handling)
- *Medium*: (a) Student describes slope in correct definitional terms (rise/run), but with no link to the context; or (b) Student describes the correct contextual variables, but with an incorrect relationship.
- *High*: Student describes the correct contextual variables with the correct relationship (total cost of each CD including shipping and handling).

Now suppose that a student types in the response, "Slope is the rise over the run," which the system recognizes as correct but having no context. The system displays feedback appropriate to the inferred (common) error.[2] For example**:** "*You've told me the correct definition of slope, but you need to explain it in terms of the problem. For example, what do the rise and run in the graph have to do with the cost of CDs and shipping and handling?*" The student then tries again, and the system uses progressive levels of feedback for scaffolded support of learning.

After each response, or some other defined interval, the system updates the relevant nodes in the student model. Thus estimates of relevant proficiencies would be updated according to the evidence model.

[2] The common errors, per item, were identified after we reviewed answers to about 500 paper and pencil tests covering all 8 proficiencies, in each of the 4 variants (i.e., graph, numeric, expression/equation, and text), and with two difficulty levels (easy and hard). After the tests were scored, incorrect responses, per item, were examined and tallied in a spreadsheet. The more frequent errors were further analyzed to infer misconceptions or procedural bugs underlying them.

The example above showcases an ETS tool called c-rater™ that can capture and analyze text input. Another ETS tool can "read" points and lines on a graph, and compare values to scoring rules (Bennett, Morley, Quardt, & Rock, 2000). Diagnostic feedback can similarly be embedded in xml files for the task, and linked to different responses. See Figure 2 for an example of graph analysis and feedback.



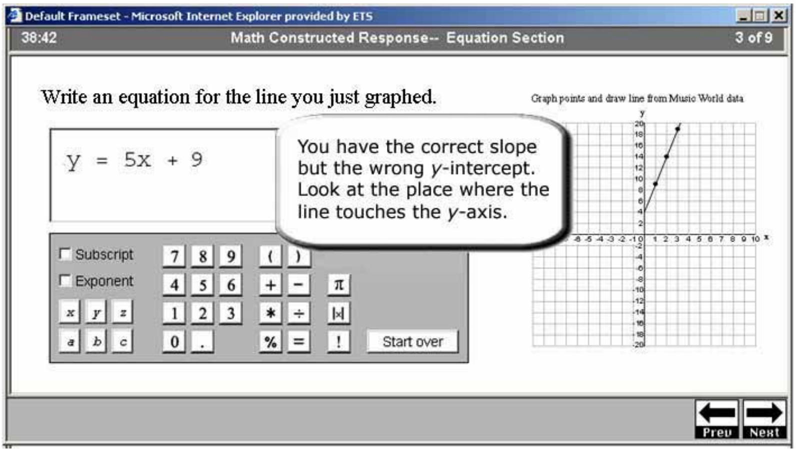
Default Frameset - Microsoft Internet Explorer provided by ETS

38:21 Math Constructed Response-- Graph Section 1 of 7

Graph points and draw line from Music World data

Good! This line shows the number of CDs and their cost.

FIGURE 2
Graph analysis with diagnostic feedback shown superimposed on the work product.



Default Frameset - Microsoft Internet Explorer provided by ETS

38:42 Math Constructed Response-- Equation Section 3 of 9

Write an equation for the line you just graphed.

y = 5x + 9

You have the correct slope but the wrong *y*-intercept. Look at the place where the line touches the *y*-axis.

Graph points and draw line from Music World data

FIGURE 3
Equation analysis with diagnostic feedback shown on the work product.

Additionally, the program evaluates the expressions and equations that a student types (see Figure 3) for mathematical accuracy/equivalence. For more information on the various automated scoring methods, see Bennett, Morley, Quardt, & Rock (2000) and Bennett, Morley, and Quardt (2000).

The various elements of MIM are based on sound principles of assessment and instruction (e.g., Kuntz, et al., 2005; Shute, 1995; Mislevy, Steinberg, & Almond, 2003). That is, we used ECD to develop the underlying proficiency model, scoring rules, and informative assessment tasks, and incorporated into MIM the three research-based features to support learning discussed in this chapter: timely diagnostic feedback, tailored content, and multiple representations of concepts. Finally, we plan to pilot test the first MIM module, employing three learning conditions: Control (classroom instruction only), Practice (classroom instruction and practice problems on relevant topics), and Treatment (classroom instruction and the MIM prototype). This will be administered to several hundred students in school districts in southern California. Of interest will be the value-added of MIM over the other two conditions in relation to student learning.

SUMMARY AND NEXT STEPS

Evidence-based learning forms the foundation of the approach proposed in this paper for the design and development of diagnostic assessments of math problem solving. Our ETS tools use a variety of evidence (e.g., performance data) as the basis for scores—as the data are aligned with rubrics. Given the range of ETS tools at our disposal, data can assume a variety of forms, such as graphs, numbers, equations, and short textual responses—scored by c-rater (for content analysis).

Problem-solving ability develops over time, and there are general (domain-independent) and specific (domain-dependent) kinds of problem solving strategies that are brought to bear on a given problem. We've chosen to focus on domain-dependent skills initially. We believe that an accurate student model flows from quality evidence, which is obtained from carefully designed assessment tasks linked to a valid proficiency model and its constituent knowledge, skills, and abilities. The local-to-global diagnostic approach is particularly powerful when coupled with sufficient practice opportunities and targeted diagnostic feedback. The next steps will involve a series of controlled evaluations to test the contributions of the various assessment and instructional elements in MIM to student learning. We also plan to determine cost-efficient ways to scale up the content to include additional mathematics concepts and skills linked to important state and national standards. The goal is to help turnaround the poor showing by U.S. students in relation to math problem solving skills as compared to their international peers.

REFERENCES

Bennett, R. E. & Bejar I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests. *Applied Psychological Measurement, 24,* 294–309.

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education, 13*, 303–322.

Corbett, A., McLaughlin, M., & Scarpinatto, K. C. (2000). Modeling student knowledge: Cognitive tutors in high school and college. *User Modeling and User-Adapted Interactions, 10*, 81-108.

Falmagne, J-C, Cosyn, E., Doignon, J-P, & Thiery, N. (2004). The assessment of knowledge, in theory and in practice. Retrieved September 22, 2005, from http://www.business.aleks.com/about/Science_Behind_ALEKS.pdf

Heffernan, N. T., & Koedinger, K. R. (2002) An intelligent tutoring system incorporating a model of an experienced human tutor. In the *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems,* 2002. Biarritz, France.

Katz, I., Lipps, A., & Trafton, J. (2002). Factors affecting difficulty in the generating examples item type (GRE Board Report No. 97-18P). Princeton: Educational Testing Service.

Koedinger, K.R., & Tabachneck, H.J.M. (1994). Two strategies are better than one: Multiple strategy use in word problem solving. Presented at the annual meeting of the *American Ed ucational Research Association,* New Orleans, LA.

Kuntz, D., Fife, J., Shute, V., Graf, E. A., Supernavage, M., Marquez, E., et al. (2005). *MIM: Mathematics Intervention Module 1.* [Unpublished computer program/prototype]. Educational Testing Service, Princeton, NJ.

Melis E. and Andrés E. (2002). About the Global Suggestion Mechanisms in Active Math. *Proceedings ITS-02 Workshop on Creating Diagnostic Assessments.*

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2000). Evidence-centered assessment design. A Submission for the NCME Award for Technical or Scientific Contributions to the Field of Educational Measurement. Retrieved November 12, 2004, from http://www.ncme.org/about/awards/mislevy.html

Mislevy, R. J., Steinberg, L.S ., and Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective, 1* (1) 3-62.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H. Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437-446). San Francisco, CA: Morgan Kaufmann.

Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437-446). San Francisco, CA: Morgan Kaufmann.

PISA Report (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. Perspective*. Retrieved July 28, 2005 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005003.

Scandura, J. M. (2005). AuthorIT: Breakthrough in Authoring Adaptive and Configurable Tutoring Systems? *Technology, Instruction, Cognition & Learning (TICL), 2*, 185-230.

Shafrir, U. (1999). Representational Competence. In I. E. Sigel (Ed.), *Theoretical Perspecitives in the Development of Representational Thought.* New Jersey: Lawrence Erlbaum Associates.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for sighted and visually-disabled students. In L. PytlikZillig, R. Bruning, and M. Bodvarsson (Eds.). *Technology-Based Education: Bringing Researchers and Practitioners Together* (pp. 169-202). Greenwich, CT: Information Age Publishing.

Tabachneck, H. J. M., Koedinger, K. R., & Nathan, M. J. (1994). Toward a theoretical account of strategy use and sense-making in mathematics problem solving. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.*_Hillsdale, NJ: Lawrence Erlbaum Associates.