Moving to the Next Generation System Design: Integrating Cognition, Assessment, and Learning

CSE Report 706

Eva L. Baker

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

February 2007

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002 and Award Number R305A050004, as administered by the Institute of Education Sciences, U.S. Department of Education, and by the Office of Naval Research, Award Number N00014-02-1-0179.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, the U.S. Department of Education, or the Office of Naval Research.

MOVING TO THE NEXT GENERATION SYSTEM DESIGN: INTEGRATING COGNITION, ASSESSMENT, AND LEARNING[1]

Eva L. Baker
National Center for Research on Evaluation, Standards, and Student Testing
University of California, Los Angeles

Abstract

This paper will describe the relationships between research on learning and its application in assessment models and operational systems. These have been topics of research at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) for more than 20 years and form a significant part of the intellectual foundation of our present research Center supported by the Institute of Education Sciences. This description serves as the context for the presentation of CRESST efforts in building the POWERSOURCE© assessment system as described in subsequent papers delivered at Session N2 of the 2006 annual meeting of the National Council on Measurement in Education.

This paper will describe the relationships between research on learning and its application in assessment models and operational systems. These have been topics of research at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) for more than 20 years and form a significant part of the intellectual foundation of our present research Center supported by the Institute of Education Sciences, as well as of many of the research studies on learning and assessment funded by the Office of Naval Research. This description is intended not to advocate a general approach (although I do), but rather to serve as the context for subsequent presentations about CRESST efforts in building the POWERSOURCE© assessment system.

Part 1. Rationale for Model-Based Assessment (MBA): Background of Our Efforts to Incorporate Learning Psychology Into Assessment Systems

Definitions of Model

CRESST R&D in assessment and learning fits two different but compatible definitions of the term "model." The first definition of model relates to the science it

[1]Paper presented at the 2006 annual meeting of the National Council on Measurement in Education (NCME). Note that the first extended section of this piece is for those who have not heard or read the many discussions of model-based assessment as practiced at CRESST. The second section can be best understood by first understanding how the model is used. The section on relating learning to assessment design and use is most directly relevant to the presentation as described in the conference program.

seeks to explore: "(3) A schematic description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further study of its characteristics: a model of generative grammar; a model of an atom; an economic model" (http://education.yahoo.com/reference/dictionary/entry/model;_ylt=AmKmF1dMF80TwgUm.ir31lOsgMMF); or, as in our case, a model of assessment based on learning. We deem CRESST's research an exercise in model building because it is intended to guide systematically the research hypotheses to be explored to understand the relationships among the model's components, as well as to assess the evidence of the model's validity for specific purposes, contexts, and students. Characteristics of the CRESST model include a set of defined components and hypothesized rules guiding their interaction. These include characteristics that constrain assessment development, for example, cognitive demands (i.e., problem solving, metacognition, domain knowledge, content representations, approximations of expertise), and the predictions that can be made about performance on sets of tasks by students possessing various types and levels of prior knowledge or exposed to differing instruction. Inferences can also be drawn about the validity of findings for specified purposes.

The second definition of model (http://education.yahoo.com/reference/dictionary/entry/model;_ylt=AmKmF1dMF80TwgUm.ir31lOsgMMF) elaborates the practical implementation of design: "(1) A small object, usually built to scale, that represents in detail another, often larger object." Considered as: "(2a) A preliminary work or construction that serves as a plan from which a final product is to be made: a clay model ready for casting; (2b) such a work or construction used in testing or perfecting a final product: a test model of a solar-powered vehicle." Here our use of model functions as a prototype or a set of properties that can be used to replicate the essential elements of a task set measuring a given domain or construct, following confirming, and preliminary, empirical evidence. These prototypes serve as componential templates that make concrete the conceptual models and augment them by specifying task instructions and scoring rules, as well as access to needed information to elicit answers. This is the definition of model that we use in practice when we are generating a range of multiple instances designed of tasks or items to estimate a student's performance at a particular time, in a series, or as an item pool. In practice, the tasks or items are composed by the combination of a particular skill set, cognitive demands, and content domain, wrapped in a task context.

That the two definitions interact should be clear. If the purpose of the scientific use of a model is to develop theoretical and empirical evidence to support its components and their relationships, then the purpose of the prototype definition of a model is to guide users in simplifying test design and development by using the prototypes to guide the creation of other instances of tasks or tests. The product of the prototype use (i.e., coherent sets of items derived from the model) gives us the tasks and tests to use in order to direct continuous attention to our scientific use of "model," that is, systematic hypothesis testing, and examination and refinement of our ideas and practices and their underlying knowledge base. So while the definitions may seem to neatly separate into "science" and "practice," in our research and development reality, it is their interaction that powers our approach.

Models for Assessment and Learning

Our use of the term "model" throughout our discussions thus should explain why we believe that scientific underpinnings of our model-based system differentiate it from others using similar nomenclature. We have used CRESST models as a conceptual and practical guide to generate multi-purpose assessment tasks, required scoring protocols and interpretations, and connected learning opportunities, for teachers in the form of professional development, and for students in order to improve their performance (Baker, 1994; Baker, Aschbacher, Niemi, & Sato, 1992/2005; Baker, Freeman, & Clayton, 1991; Baker et al., 1996; O'Neil, Chung, & Brown, 1997) and collaborative problem solving (Baker & Mayer, 1999). For example, one attribute of our model is that scoring schemes are developed using expert-novice comparisons (see Chi, Glaser, & Farr, 1988, for multiple examples). The components of our model can be empirically tested to determine whether they contribute to hypothesized relationships among variables (e.g., score differences between instructed and uninstructed students). They may also be used to assess whether our tasks and results contrast with those of tests prepared using other approaches. These comparisons are made by looking at performance of students using both examinations and determining whether predictive validity or equity advantages attended comparative methods (Martínez-Fernández & Goldschmidt, in press). One interesting study in this regard was completed by Goldschmidt and Martínez-Fernández (in press) using the California High School Exit Examination as a criterion. Even though items may appear to be similar, visual inspection and review will not do the trick, because despite surface features, it is our contention that

items created with the CRESST model in mind will generate different consequences and different validity inferences.

Model Components

The major components of the CRESST model are (a) domain-independent and domain-specific cognitive demands of the task or test, (b) criteria to judge performance derived from expert performance, and (c) detailed representation of the content map that shows the topology of the subject matter (ontology; see Figure 1). For formative purposes, we also need to attend to the nexuses depicting individual and shared dependencies of learning. These may be identified through empirical means, but are typically verified in a form of cognitive task analysis (Clark, 2004; Ericsson, 1996) or using other operational representations of dependencies (Gagné, 1985). The last part of the model addresses utility: the practical matters involved in constructing the tasks or items in a form that will be realistic and permit the economic regeneration of items and tasks (by the reuse of certain frames or strategies), making decisions relevant to time constraints for development or administration, and avoiding the inadvertent introduction of construct-irrelevant components, particularly for students of different backgrounds and sensitivity to the political context of administration.

Embedded Research

It is our goal that all serious decisions, whether for scientific exploration or practical application, should be made on the basis of empirical evidence, and revisions or changes should be made in a principled manner (a goal we try to achieve, unless occasionally thwarted by implacable user constraints). One controversial feature of our models is that they begin with syntheses of research that derive from evidence in more than one domain. So the initial components of the model (e.g., cognitive demands) are conceived in a domain-independent form, with the expectation that attributes of our assessments can cross topic domains. For example, problem identification or data conflict reduction strategies can be used in a variety of very different content areas and settings. Paired with these domain-independent components are attributes derived from subject-matter specific domains (e.g., rules for the use of figurative language). By embedding our "domain-independent principles," such as attributes of problem solving (see O'Neil, Chung, & Brown, 1997; or Klein, 1989), in the context of a subject matter, and using principles or strategies unique to subject matter, we believe that we will be

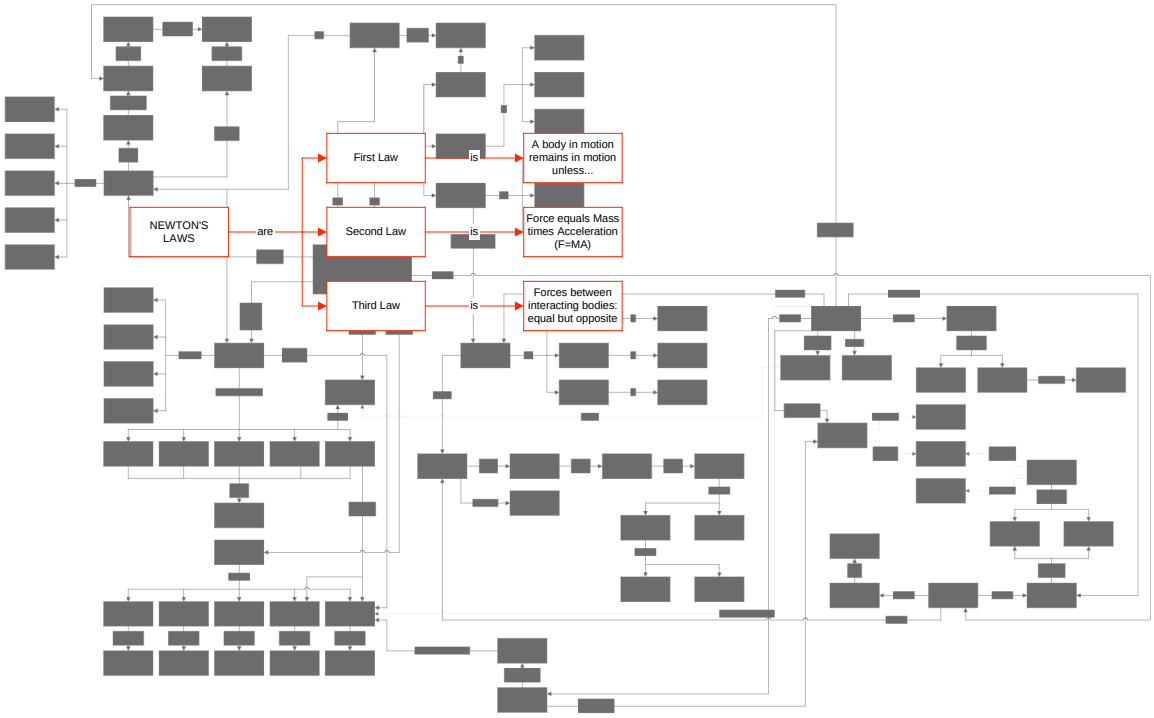

Figure 1. Physics ontology.

instrumentally supporting generalization and transfer. Of course, there are not sharp divisions between domain-independent and domain-specific strategies, but rather a continuum of knowledge and strategies, anchored at one end by those that have only limited topic or subject-matter applications and at the other end by those that have a much broader, general use. For example, in writing performance, there are some organizational rules guiding the preparation of persuasive essays. These rules, such as *refute negative arguments*, could be applied across a wide number of content areas that a student may be asked, or may choose, to address. It may not matter much for the effectiveness of these rhetorical rules whether the topic is avoiding trans-fats (for instance, arguing for the role of pharmaceutical development) or considering ideal savings plans for middle-aged couples. The domain-independent rules (e.g., persuasive strategies) apply and are in fact reusable, even though the specifics of each argument would be embedded either in pharmaceutical content knowledge or in various equity and cash sequestering plans.

This approach has demonstrable economic implications for test development. Because our assessment models hinge partially on attributes that are largely domain

independent, elements of the model, operationalized as a template, may be adapted to other content, topics, or users. In addition to domain-specific research findings relevant to learning, the second major element of our work is the detailed representation of the content domain as an ontology or connected map. It is the flesh that fills out the bones of our models. These ontologies present both content elements to be learned and hierarchies or lateral relationships. The ontology presents a structure of domain-specific knowledge, such as declarative, procedural and strategic knowledge, to be used to guide both assessment design and instructional practice. Ontologies used in this form have their source in computer science. In our experience these ontologies can fully represent a subject matter, at a grade level or overall, or in contrast, focus explicitly on a subdomain for the design of testing and instruction of identified components. Examples of areas in which CRESST researchers have recently created maps include mathematics, college-level engineering, the force and motion topics in middle school physics, and procedural skills in marksmanship.

Model Rationale Summary

To summarize the argument about model-based assessment, we expect that the balance between domain-independent and domain-specific knowledge and components will vary with the type of learning intended. Nonetheless, we believe that it is a useful perspective to guide research on assessment. It may suggest that in every test or task there is some aspect that can be explicitly evoked from student background knowledge or existing schema. That alone will avoid *de novo* design and encourage the use of assessment designers' store of knowledge about learning. In fact, the core premise of our current version of model-based assessment is that the major function of assessment practice is to illuminate and support students' learning, a holdover from the performance assessment days. In our current view, student learning is not only directed to the desired outcome measures (i.e., external accountability measures like state assessments), but also encompasses variations in setting and constraints that should enable students to demonstrate their proficiency on tasks calling for generalization and transfer, beyond, for instance, the form of the standards-based test that is used. For systems that have a formative purpose, the model must also reliably identify weaknesses at the conceptual or skill (not item analysis) level. The reliance of model-based assessment on the science underlying the relationships among cognitive and content components of assessment and instruction and the procedures for replicating such assessments in regular practice

give it potential strength and certainly an ambitious, rich research agenda. Although much of our model discussion applies to the design work preceding actual test administration, our interests and research also include the consequences of particular assessment results and how to improve less-than-acceptable performance. So integrated with models constraining assessment design are connected structures about professional development, the ways in which instruction might be variously accomplished, the role of feedback, and developing the students' own responsibilities as the managers of their own learning. *It should be underscored that the instructional decisions made by teachers should be related intimately to the model and ontology generating the assessment, rather than to the assessment items themselves.* Our research interest is directed to a set of connected hypotheses about characteristics of instruction, student learning, aspects of subtask performance, background characteristics, and performance on targeted assessments and on transfer measures. Because we claim our designs serve multiple assessment purposes (e.g., certification and diagnosis), we share substantial problem space with researchers in formative assessment, many of whom may make very different assumptions about item design.

Part 2. The Role of Cognitive and Other Learning Research in the Design of the POWERSOURCE© System

Rather than discuss the warrants for classroom or formative assessment, at this point I wish to highlight the cognitively oriented features displayed in our newly funded assessment Center's research and describe some of the design elements underlying the student measures, the professional development, and strategies to support learning and instruction throughout the year that will inform the POWERSOURCE© system. The first iteration of this version is on the topics of pre-algebra and algebra for Grades 6–8.

As an aside, it is more accurate to describe our research base as eclectic, evolving from research on learning writ large, rather than from cognitive psychology in particular. In any case, we especially wish to avoid potentially traumatic contentions about which features represent either new behaviorism, old behaviorism, or the situated, social, or structural interpretations of cognition. Instead of a coherent, single theory, we have selected from the research the most promising findings from scientific knowledge derived from studies of learning and instruction to use as the foundation for the assessment system in our current project.

Learning Sources for POWERSOURCE© Design

This section explicitly addresses the scientific features underpinning assessment design, its classroom use, its attendant professional development of teachers, and a set of strong supports for teachers and students to extend and intensify the treatment. There are 10 such design features, supported by illustrative research guidance (see Table 1). Note that items 1 and 2 have been described above in the discussion of CRESST models. I also believe that cognitive task analysis and

Table 1

Research-Based Learning Knowledge as Design Features of POWERSOURCE©

Design feature	Research warrants
1. Synthesis of domain-independent models	Baker, 2003; Baker & Mayer, 1999; Bransford & Johnson, 1972; Niemi, 1996; O'Neil, Chung, & Brown, 1997
2. Ontology (detailed structural relationship among content components)	Bruner, 1964; Chung, Delacruz, & Bewley, 2004; Niemi, 1996; Vendlinski, Niemi, Wang, & Monempour, 2005
3. Cognitive and content dependencies; domain-specific learning	Ericsson & Simon, 1998; Gagné, 1985; Klein, 1989; Klein, Chung, Osmundson, Herl, & O'Neil, 2002; Zachary, Ryder, & Hicinbothom, 1998
4. Expert-novice criteria for performance	Baker, 1997; Baker, Freeman, & Clayton, 1991; Chi, Glaser, & Farr, 1988; Ericsson, 1996
5. Schema development	
– to reduce cognitive load	Mayer, 2003; Sweller, 1999
– to heighten transfer and generalization	Bassok & Holyoak, 1989; Bjork & Richardson-Klavehn, 1989; Mayer, 2003; Mayer & Wittrock, 1996; Sweller, 1999
– worked examples for teachers and students	Mayer, 2003, Merrill, in press; Sweller, 1999
6. Explanation to internalize learning	Chi, 2000; Niemi, 1996
7. Feedback	Nyquist, 2003
8. Language characteristics	Abedi, Hofstetter, & Lord, 2004
9. Scaffolding using performance aids	Ausubel, 1972; Bruner, 1964; Collins, Brown, & Newman, 1989; Guthrie, 1959
10. Motivation and engagement	
– narrative support	The Cognition and Technology Group at Vanderbilt, 1997; Gudmundsdottir, 1991, p. 209; Merrill, in press; Mor & Noss, 2004

domain-specific models (item 3) are well known to the audience. It is likely that most of you are also familiar with the expert-novice literature (item 4). Here I will only reiterate that we use these contrasts as ways of assuring that the tasks are sequenced appropriately and to identify from expert or expert-like performance the criteria that will be applied through the training of raters. Item 7, feedback, has a deep research history that I need not explore here, and item 8, while of extraordinary importance, here is described as research that applies linguistic rules to protect against assessment/student group interactions from construct (domain) irrelevant variance; that is, basing performance inferences on results that are less about the content-related performance, and rather, that reflect complexities in the language choices in the assessment.

This leaves items 5–schema development, 6–explanation, 9–scaffolding using performance aids, and 10–motivation and engagement through the use of connected narrative. Items 5, 6, and 9 all support the development of transfer and generalization, attributes that will be measured as part of the dependent measures used in our experiment.

We have recently adopted a number of approaches intended to facilitate rapid acquisition of skills and to support transfer and generalization, through the use of schema development. It is our intention that by focusing on a handful of "big ideas" in the field over the sixth, seventh, and eighth grades, we will induce schema related to the core understandings needed for success in algebra (see Figure 2). A strategy we are testing to attain these goals involves the use of worked examples (Mayer, 2003; Merrill, in press; Sweller, 1999). Sweller and Mayer invoke the theory of cognitive load in support of worked examples. They suggest that complex tasks, procedural or problem solving, may be best acquired and rapidly accessed when they are taught as coherent worked examples rather than by learning to criterion individual sets of steps. Chi (2000) also supports this approach to schema development. The logic is that when students encounter a problem or situation with similar properties and can identify the problem type or principles that underlie its solution, the "worked example" is accessed, and students will be able to adapt their previously learned schema to new situational requirements. Worked examples may be pursued individually or in groups. This approach is thought to conserve working memory because, in contrast to step-by-step recall of the elements and the sequence needed to make an appropriate response to a complex problem, the learner retrieves a more integrated unit. Sweller and his colleagues have developed studies

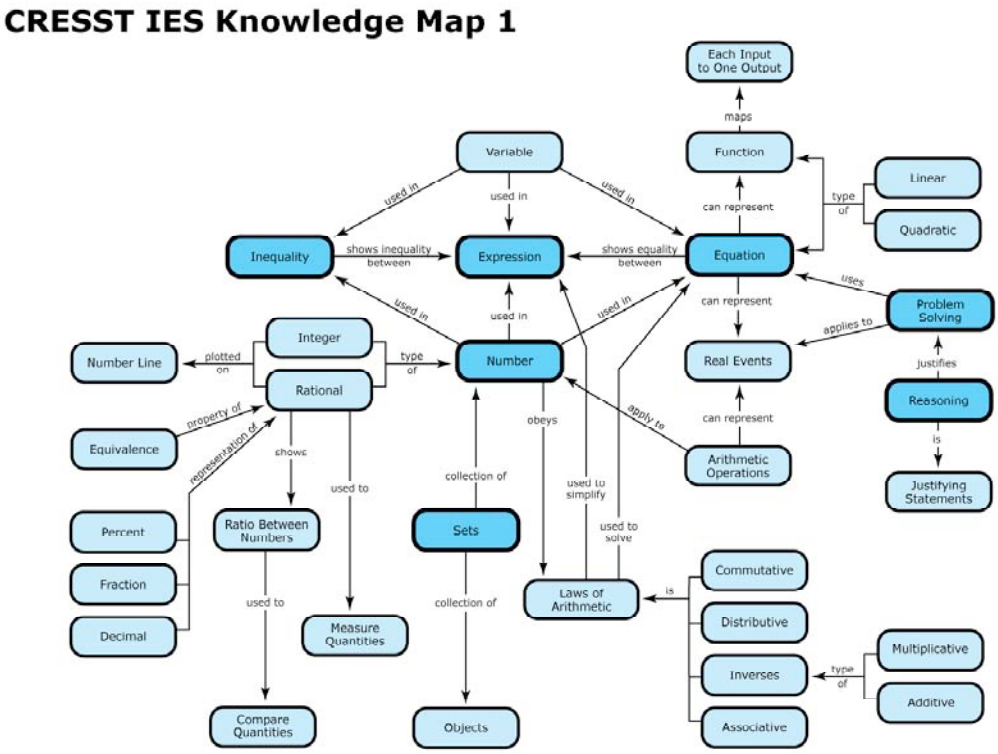

Figure 2. Algebra mapper.

suggesting the transfer and generalization impact of this approach. While worked examples have been generally used as approaches to improve instruction, we have chosen to use them in part as an additional template in our model-based assessment development. In the tasks themselves, students will be asked to complete partial examples of complex problems. Because students will be given opportunity within the assessments to address varying constraints using the same schema, we expect to find transfer and generalization to a broader domain of problem sets than might be obtained by traditional benchmark testing or examination preparation strategies. An additional research-based strategy (item 6) to augment the worked-example approach requires students to write explanations of *why* they took particular courses of action, and not simply specify in language w*hat* they did. It has been documented by research in CRESST's work that explanation skills can be taught for subject matter topics, that they can be reliably scored and that they predict deeper understanding of content. Arguments on behalf of verbal explanation also support the ability to identify problem contexts and, in fact, can help scaffold learning.

Scaffolding (item 9) is a commonly understood idea, but we intend to use it in a relatively innovative way in this work, as it will take the form of concrete (or electronic) performance aids. Performance aids have a long history (see Ausubel, 1978; Bruner, 1964; Guthrie, 1959) and are an operating technology in much of the business and military training world. We have them on our photocopiers and cell phones, and they are often used to support procedural performance. We cite Collins, Brown, and Newman (1989) to justify their use in problem solving:

> Scaffolding can be applied to different aspects of a problem-solving process, for example, to management and control of the problem solving or to the subprocesses that are required to carry out the task. Global before local skills means that in the sequencing of lessons there is a bias toward supporting the lower level or composite skills that students must put together to carry out a complex task. In algebra, for example, students may be relieved of having to carry out low-level computations in which they lack skill to concentrate on the higher order reasoning and strategies required to solve an interesting problem. (p. 485)

We have begun new work in developing the types of performance aids that teachers and students might use. Just as managers and intellectual workers are not expected to have every detailed piece of knowledge memorized, we also intend to supply students and teachers with task performance aids. The task performance aids we intend to use will provide scaffolding in three specific ways: in laminate card form (a teacher and a student version), in poster and other displays of the task ontology, and in Web sites. The Web sites will include the maps (as we call ontologies for students) and worked examples. In addition, the Web sites will be populated with a bagful of micro-instructional interventions (we alternately call them "parcels," "vitamins," or "bursts"). These will be 1- to 2-minute instructional reminders for student use, directed either to common errors or missing prior knowledge. We expect that these "parcels," "bursts," or "vitamins" will provide instructional supports that can be used on the fly, augmenting adopted texts and materials, to help support acquisition, retention, and transfer (see Figure 3). Although our primary focus is student learning, we will be using similar strategies to assist the teacher. For teachers we will also use worked examples designed for teachers, to give them help when either their existing repertoire of options is limited, or as a delicate way of approaching shortfalls in domain knowledge or pedagogy. In our program, we plan to use them in professional development (to extend the treatment time available) or, just as likely, when the instructional time available for remediation or reteaching is unscheduled or far too brief. Task performance aids

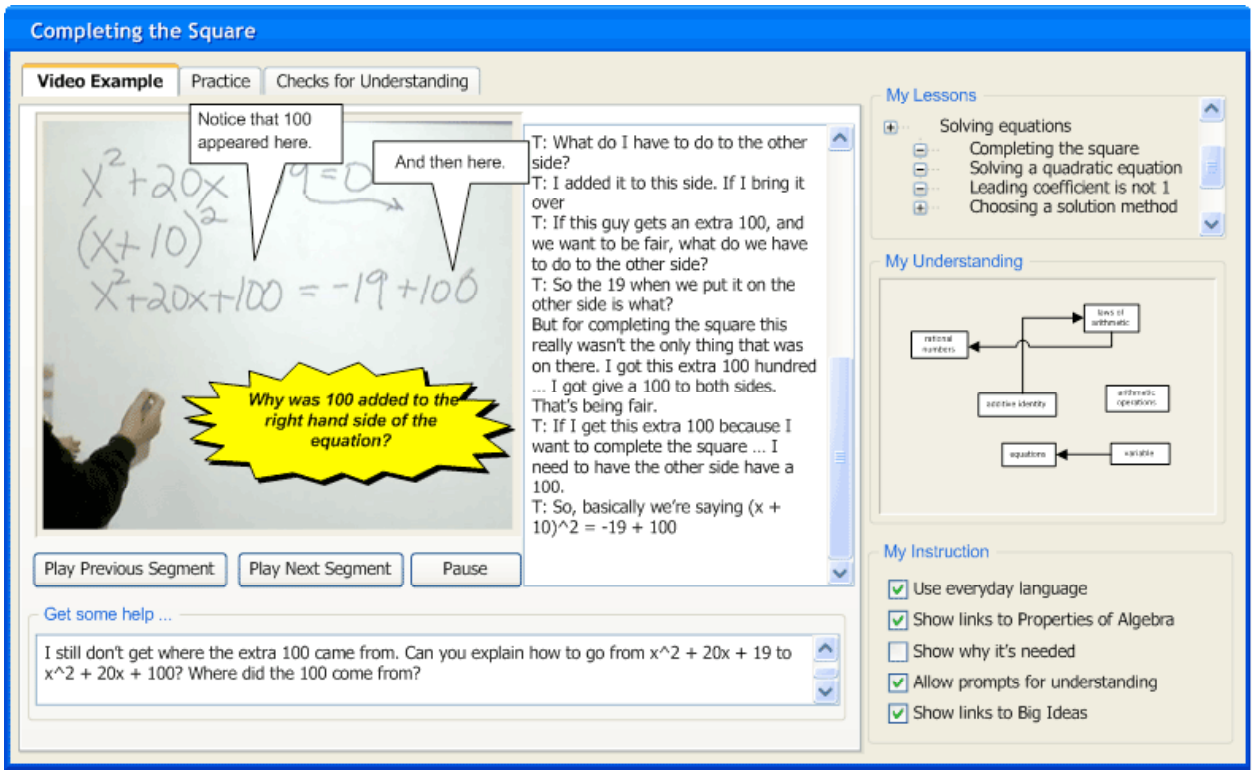

Figure 3. Computer burst concept GUI.

will be structured so that they remind students (and teachers) of where they are in the ontology, what the big principles (schema-related ideas) are, how worked examples look, how to solve simple problems or more complex word problems, and to how create explanations. They will also be available (and potentially experimentally compared) for students and teachers for scoring or self-assessing their work including a variety of worked examples and maps of the content domains for students and teachers.

Item 10, motivation and engagement, has many interpretations. We are experimenting this year with the use of coherent narrative structure (The Cognition and Technology Group at Vanderbilt, 1997; Mor & Noss, 2004) to provide linkage and coherence for the POWERSOURCE© assessment system. Stories will be framed around the problems; to advance or find the answers, students will need to solve problems. As in the Jasper work, some of the problems themselves will be story related. We intend to use a coherent look and feel for our assessments, scoring schemes, task performance aids, Web site and other ancillary materials. We are engaging professional writers with two sets of experiences: (a) prior success in

working with middle and high schools, and (b) skills in the anime or Manga genres of graphic novellas. We are testing story lines and characters for interest and hope to develop additional funded studies that look at the effect of narratives on learning, transfer, generalization, interest, motivation, reading skills, and of course, distraction.

So despite how it may sound, we are not naïve enough to think that a few well-designed tests will compete easily with district-required exercises that carry with them very short term and strong sanctions. Our investment in these aspects of assessment design taken from instruction, self-motivated learning, and the commercial sector, highlights the evolution of research-embedded, model-based assessment as markedly different from our past—assessment-only—approach. If we are successful, these strategies can be inserted into the ongoing curriculum (slipped in, actually), in concert with or without benchmark examinations, and will take only a relatively few minutes spent regularly across a 3-year period. We expect that students in our experimental group who have experienced the full treatment will show differential growth on examinations and, more importantly, will dramatically outperform the control group on transfer tasks, including items selected from examinations (Key Stage 3, in England; PISA; and other state assessments).

References

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*, 1-28.

Ausubel, D. P. (1972). In defense of advance organizers: A reply to the critics. *Review of Educational Research*, *48*, 251-257.

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist*, *29*, 97-106.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice*, *36*, 247-254.

Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues & Practice*, *22*(2), 13-17.

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992/2005). *CRESST performance assessment models: Assessing content area explanations* (CSE Rep. No. 652). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior*, *15*, 269-282.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muñoz, Z., Staley, L., & Linn, R. L. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bassok, M., & Holyoak, K. J. (1989). Transfer of domain-specific problem solving procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 522-533.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717-726.

Bruner, J. S. (1964). Some theorems on instruction illustrated with reference to mathematics. In E. R. Hilgard (Ed.), *Theories of learning and instruction* (National Society for the Study of Education Yearbook, Vol. 63, Part 1, pp. 306-335). Chicago: National Society for the Study of Education. Distributed by the University of Chicago Press.

Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in instruction psychology* (Vol. 5, pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, R. E. (2004). What works in distance learning: Motivation strategies. In H. O'Neil (Ed.), *What works in distance learning: Guidelines* (pp. 89-110). Greenwich, CT: Information Age Publishers.

The Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, instruction, assessment, and professional development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A. (Ed.). (1996). *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, & Activity, 5*, 178-186.

Fletcher, J. D., & Morrison, J. E. (in press). Representing cognition in games and simulations. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gagné, R. M. (1985). *The conditions of learning and theory of instruction* (4th ed.). New York: Holt, Rinehart & Winston

Goldschmidt, P., & Martínez-Fernández, J.-F. (in press). The relationship among measures as empirical evidence of validity: Performance assignments, SAT-9, and high school exit exam performance incorporating effects of school context. *Educational Assessment*.

Gudmundsdottir, S. (1991). Story-maker, story-teller: Narrative structures in curriculum. *Journal of Curriculum Studies, 23*, 207-218.

Guthrie, E. R. (1959). Association by contiguity. In S. Koch (Ed.), *Psychology: A study of a science, Vol. 2. General systematic formulations, learning, and special processes* (pp. 158-195). New York: McGraw-Hill.

Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., Herl, H. E., & O'Neil, H. F., Jr. (2002). *The validity of knowledge mapping as a measure of elementary students'*

scientific understanding (CSE Tech. Rep. No. 557). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Klein, G. A. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.), *Advances in man-machine research* (Vol. 5, pp. 47-92). Greenwich, CT: JAI Press, Inc.

Martínez-Fernández, J.-F., & Goldschmidt, P. (in press). Comparing student and teacher reports on opportunity to learn and their relationship to student achievement. *Educational Assessment*.

Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Merrill Prentice Hall.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47-62). New York: Macmillian Library Reference USA, Simon & Schuster Macmillan.

Merrill, M. D. (in press). First principles of instruction: Instructional design In C. M. Reigeluth & A. Carr (Eds.), *Instructional design theories and models III*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mor, Y., & Noss, R. (2004). *Towards a narrative-oriented framework for designing mathematical learning*. Paper presented at the CSCL SIG First Symposium, Lausanne, Switzerland.

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representations, problem solutions, justifications, and explanations. *Journal of Educational Research*, *89*, 351-363.

Nyquist, J. B. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis*. Unpublished master's thesis, Vanderbilt University, Nashville, TN.

O'Neil, H. F., Jr., Chung, G. K., & Brown, R. S. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil (Ed.), *Workplace readiness: Competencies and assessment* (pp. 411-452). Mahwah, NJ: Lawrence Erlbaum Associates.

Sears, P. S., & Hilgard, E. R. (1964). The teacher's role in the motivation of the learner. In E. R. Hilgard (Ed.), *Theories of learning and instruction* (National Society for the Study of Education Yearbook (Vol. 63, Part 1, pp. 182-209). Chicago: National Society for the Study of Education. Distributed by the University of Chicago Press.

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.

Vendlinski, T. P., Niemi, D. M., Wang, J., & Monempour, S. (2005). Improving formative assessment practice with educational information technology. In F.

Malpica, F. Welsch, A. Tremante, & J. Lawler (Eds.), *The 3rd International Conference on Education and Information Systems: Technologies and Applications: Vol. 1* (pp. 361-366). Orlando, FL: International Institute of Informatics and Systematics.

Zachary, W. W., Ryder, J. M., & Hicinbothom, J. H. (1998). Cognitive task analysis and modeling of decision-making in complex environments. In J. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 315-344). Washington, DC: American Psychological Association.