Tech., Inst., Cognition and Learning, Vol. 2, pp.1-18
Reprints available directly from the publisher
Photocopying permitted by license only

Towards Automating ECD-Based Diagnostic Assessments

VALERIE J. SHUTE

Educational Testing Service, Rosedale Road, MS 13E, Princeton, NJ 08541 USA

In this paper, I briefly present current problems in the U.S. educational system, overview the evidence-centered design (ECD) approach for developing assessments, and describe how one can address the problems by using the ECD approach for designing online diagnostic assessments in support of learning. I also describe a method for modifying an existing knowledge elicitation tool that can assist in obtaining the necessary data for each of the student and evidence models embodied within the ECD framework.

Keywords: Assessment objects, Bayesian networks, content model, diagnostic assessment, evidence-centered design, evidence model, formative assessment, knowledge elicitation, knowledge structures, learning objects, proficiency model, student model, summative assessment, task model.

INTRODUCTION

The issues and research described in this paper generally focus on enhancing K-12 education by automating the development of valid diagnostic assessments that may be used directly in support of teaching and learning. This is expected to lay the groundwork for and point towards more customized adaptive solutions that adjust to the specific needs of each learner.

I begin by briefly presenting current problems in the U.S. educational system. This is followed by an overview of the evidence-centered design (ECD) approach for assessment. The point of ECD is to engender the design

Any opinions expressed in this article are those of the author and not necessarily of Educational Testing Service.

*Corresponding author: vshute@ets.org

of valid diagnostic assessments that support learning. Following the ECD overview, I propose an idea for modifying an existing knowledge elicitation tool that can assist in obtaining the necessary data to populate the proficiency and evidence models embodied within the ECD framework. This is followed by a discussion on learning objects and their role in making these diagnostic assessments come alive. I conclude with thoughts on the near future of education.

What Are The Problems?

> *Ideally, an assessment comprises an important event in the learning process, part of reflection and understanding of progress. In reality, student assessments are used to determine placement, promotion, graduation, or retention.*

For the past couple of decades, we have witnessed a flurry of activities in the U.S.—from local to national levels—focused on improving educational achievement, opportunity, and equity. Some of the avenues pursued towards these goals include the specification of disciplinary standards, development of new instructional materials, and formation of educational partnerships and policies to create, implement, and support changes in teaching and learning. Yet despite these efforts, recent reports of the poor performance of U.S. students, relative to national and international benchmarks, call attention to the gulf between the anticipated outcomes of educational innovation and those actually reached. For example, the TIMSS results indicate that U.S. fourth graders perform above the international average in mathematics (National Center for Educational Statistics, 2001a; Office of Educational Research and Improvement, 2001). However, U.S. eighth-grade students perform at or below the international average; and by the end of high school, U.S. students perform far below the international average (International Study Center at Boston College, 1998; Office of Educational Research and Improvement, 2001). These findings have given rise to several different factions calling for diverse solutions, such as: (a) return to basics, (b) increase emphasis on conceptual understanding, and (c) require greater accountability for teachers and students.

The decision of whether schools should concentrate on basic skills versus conceptual understanding represents a fruitless dichotomy. That is, it shouldn't be construed as a disjunction, but a conjunction, where the range of knowledge and skills are instructed and assessed in a systematic and

appropriate manner. The issue of accountability, though, remains a legitimate problem. For any new policy or educational innovation to have a chance at succeeding, teachers, students, and schools must assume responsibility for their roles and actions related to teaching, learning, and supporting the learning environment, respectively. With regard to accountability measures, two important questions are: (1) how will they be fairly assessed, and (2) on what basis will rewards and sanctions be distributed? There obviously needs to be a very good metric by which to judge teachers, students, and schools. Standardized testing has thus been advocated to benchmark teachers, students, and schools as a mechanism for monitoring individual performance as well as instructional changes.

Tests alone can't improve education

One of the most prominent advocates of standardized testing is President Bush, who has placed reform of the public education system as a very high priority of his administration. Accordingly, he's calling for annual testing in grades 3-8 to get lagging or failing schools on track, saying student proficiency in math, reading, and soon science must be measured so parents will know whether their children are advancing, and administrators will know if their schools are meeting local standards. However, as Snow & Jones (2001) point out, tests alone cannot enhance educational outcomes. Rather, tests can guide improvement—presuming they are valid and reliable—if they motivate adjustments to the educational system (i.e., provide the basis for bolstering curricula, insure support for struggling students, guide professional development opportunities, and distribute limited resources fairly). "If we take no action to improve teaching and learning, we will just be using children as 'extras' in a high profile political drama while undermining the social and economic prospects of the nation in the process" (Kurt Landgraf, 2001, p. 4).

Current testing formats/procedures do not meet new criteria

Assessing large numbers of learners, via paper-and-pencil, forced-choice format, is increasingly viewed as substandard because of the overemphasis on specific student skills with a lack of attention to proficiencies, particularly those representing complex cognitive skills. The problem does not reside with the assessors as much as with those who use the measures to form simplistic and occasionally wrong conclusions, and then use those data to develop policies and/or disperse funding.

Consequently, we need to accurately *diagnose* students for particular

strengths and weaknesses across an array of knowledge and skills. This idea is similar to how pediatricians assess infants for developmental progress, comparing each with some expected or normal developmental sequence. The assessment can indicate if the individual is on track, delayed, or ahead of what is typical of the age group. It's also possible to ascertain changes in relation to the child him/herself. For instance, a particular child may reside in the bottom percentile in terms of weight and height for her age group, but show personal growth along both dimensions across the past 5-6 months. Thus, diagnostic data can be used for both norm- and criterion-referenced assessments—indicating how one is doing in relation to others, and in relation to one's own progress on a criterion or along some dimension.

In addition to the paucity of developmental information available from traditional large-scale tests, there is the very real problem associated with the long delay between when students take these tests and teachers receive the scores. That is, by the time the results of high-stakes accountability tests are disseminated, it is often too late to affect change in the classroom to address any problems. A computerized diagnostic assessment system, such as the one that will be described herein, can effectively lower the lag time by allowing teachers to recognize student difficulties immediately, and do something about them well in advance of the high-stakes tests.

Summary of Problems

The general push for increased accountability encourages an educational reform program that could possibly fail—i.e., frequent, end-of-year high-stakes testing, where the achievement gap (i.e., the difference in school performance relating to race or ethnicity) continues to exist; or worse, enlarge. There will soon be a very high demand for states to develop many new large-scale assessments in line with the new Bush policy. However one complication is that large-scale assessment has come under renewed attack. According to Bennett (2001), the issues driving the criticisms include: an outmoded basis for test design, a mismatch with curriculum, differential performance of population groups, a lack of information to help individuals improve, and inefficiency. Additional concerns center on the quality and utility of test scores, as well as the delay between students taking tests and teachers receiving the scores.

In short, the main problems in education today include the following: (a) U.S. students, starting at about middle school, need to improve their learning, especially in the areas of reading, math, and science; (b) instruction needs to focus less on low-level learning (e.g., memorization of facts) and

more on supporting higher-level problem solving skills and understanding; and (c) classroom assessments should focus less on summative tests, and more on formative, diagnostic assessments for adequate and timely learning opportunities.

What Are Some Solutions?

> *The nature of the construct being assessed should guide the selection or construction of relevant tasks, as well as the rational development of construct-based scoring criteria and rubrics.* Sam Messick, 1992.

Unified assessment-design framework and associated tools

According to Mislevy, et al. (2000), "An assessment that pushes the frontiers of psychology, technology, statistics, and a substantive domain all at once cannot succeed unless all are incorporated into a coherent design from the very beginning of the work." (p. 25). Towards this end, scientists at ETS and elsewhere have begun to pave inroads. For example, Mislevy, et al. have developed an approach called evidence-centered design (ECD) that defines a framework allowing a test developer to: (a) define the *claims* to be made about the students (i.e., the knowledge, skills, and abilities to be measured), (b) establish what constitutes valid *evidence* of the claim (i.e., student performance data demonstrating varying levels of mastery), and (c) determine the *tasks* that will elicit that evidence. This research has spawned a wide and varied group of models that are being used to design, implement, deliver, and maintain assessments. Furthermore, the models themselves have spawned a growing collection of distinctive tools. In a later section of this paper, I will describe a plan to modify an existing knowledge elicitation tool that can be used to capture both claims and evidences from subject matter experts for inclusion in proficiency and evidence models, respectively.

The ECD framework is presented in Figure 1. The assessment design process conceptually flows left-to-right, although in practice it is less linear and more iterative. To illustrate, proficiencies and claims are first delineated, followed by the linking of evidences to the claims. Finally, task requirements are specified that serve to elicit those evidences that in turn, link back through the evidence model, as shown in the figure, to the underlying proficiencies. Diagnosis flows in the opposite direction. Once a student has responded to an assessment task, evidence is identified and scored, and the

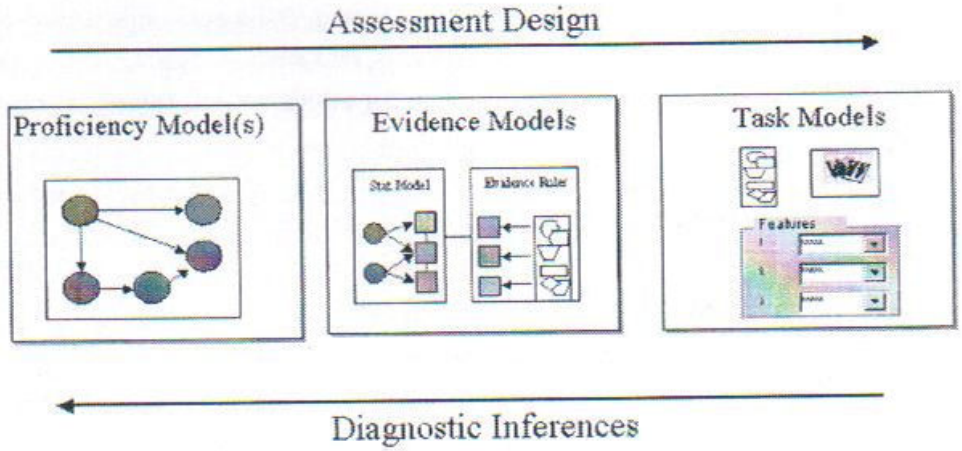

FIGURE 1
Simplified Outline of ECD framework (from Mislevy, Steinberg, & Almond, 2000).

proficiency model is updated with this new information. Diagnoses are a function of the relative, updated proficiency model values.

Proficiency Model: What to infer? The proficiency model represents a general map of the knowledge/skill terrain in a given subject area (e.g., 8th grade algebra, 4th grade reading comprehension), complete with mountains and valleys—or difficult and easy topics, and all things in between. Any such map may be instantiated (overlayed) with actual student data, which is commonly referred to as a student model.

Figure 2 shows an example of a proficiency model representing a subset of topics from 8th grade algebra, relating to sequences. Note that there are different ways to represent these nodes. For instance, one could represent higher-level nodes for understanding tables, inducing rules, and so forth, with sub-components reflecting arithmetic, geometric and other recursive problems. The breakdown employed here reflects input from middle-school teachers, who indicated that they preferred teaching these three areas separately, typically beginning with the easier arithmetic sequences.

In order to accurately infer and model proficiency levels, per student, per topic, one can use probabilistic estimates of knowledge/skill level (or mastery status) that arise from either a Bayesian network or regression equations (e.g., Mislevy, Almond, Yan, and Steinberg, 2000; Shute, 1995). Alternatively, one may use different kinds of reasoning for different processes in a larger system, such as Bayes nets for managing uncertainty about proficiency model variables (i.e., evidence accumulation), neural nets for evaluating complex strings of extracted features from a work product (evidence identification), and rule-

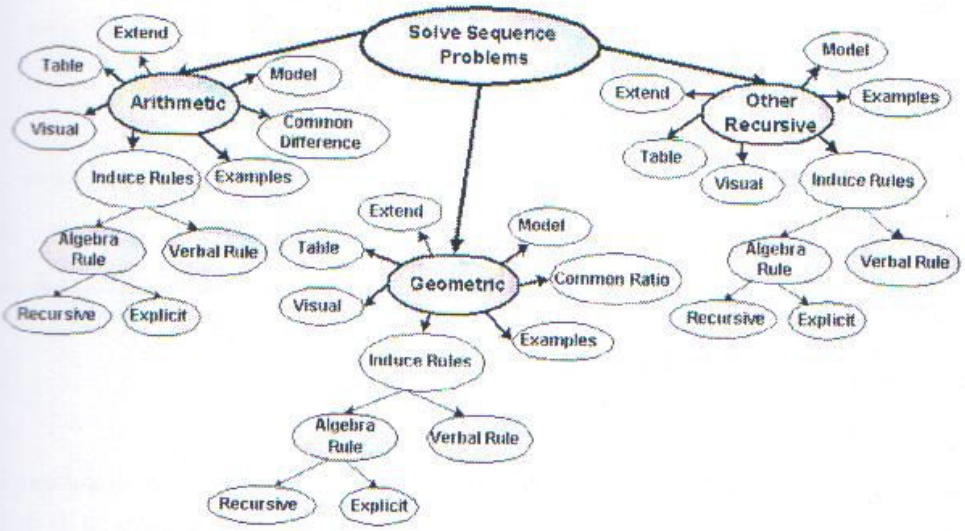

FIGURE 2
Proficiency Model for a Part of 8th Grade Algebra (Graf, 2003).

based systems for extracting features and deciding what to do next based on the current state of the proficiency model (activity selection).

Evidence Model: What data must be collected for the inferences? Individual responses to carefully-crafted items and tasks, as well as patterns of responses, serve as the primary basis for evidence of proficiencies. This information may be culled directly from the students' online behaviors as they interact with and complete items within an assessment task.

To highlight the difference between assessment and diagnosis as well as provide examples of evidence for particular cognitive skills, consider an assessment of second graders' knowledge/skill in double-digit addition. For the two problems shown in Figure 3, suppose that student **A** answered 61 and 83. The interpretation of these constructed responses (with some degree of confidence) would be that she understood, and could successfully apply, the "carrying procedure." Now view the other three students' responses. Simple assessment measures typically do not differentiate among incorrect solutions. Consequently, ensuing remediation, if any, would require all three students to re-do the specific unit of instruction.

An obvious problem with this approach is that there is often little difference between the remedial and original instruction. In other words, there is no good "fix" to the students' problem(s). A more sensitive (or *intelligent*) response by the system would be to diagnose/classify **B**'s answer as a failure to carry a one to the tens column; **C**'s answer as the incorrect adding of

	Problem 1 22 + 39 = ___	Problem 2 46 + 37 = ___
Student A	61	83
Student B	51	73
Student C	161	203
Student D	61	85

FIGURE 3
Two math problems and four students' solutions.

the ones column result (11 and 13) to the tens column; and **D**'s as a probable computational error in the second problem (mistakenly adding 6 + 7 = 15 instead of 13). Remediation could specifically address each of the three qualitatively different errors.

However, instead of modeling the plethora of potential "buggy solutions" (which can get very unwieldy, very quickly), an attractive alternative approach is one that has been successfully employed by researchers using Diagnoser (e.g., Graf, Bassok, Hunt, & Minstrell, this volume; Minstrell, 2000). For any given problem, there may be a range of answers that reflect different kinds of common conceptions. These options (erroneous and correct) each have a "rationale" associated with it that the learner chooses as his/her reason for responding with the answer he/she did. This attempts to get at the learner's current understanding or misunderstanding. It is possible to build on the idea to elaborate different kinds of tasks—beyond multiple choice that define *Diagnoser's* format—and research issues that were previously not addressed (e.g., systematically identifying difficulty factors). See Figures 4 and 5 for an illustration of a diagnostic assessment relating to finding the probability of an event that does not have equally likely outcomes, based on Brian Greer's example generated for Hunt & Minstrell's *Diagnoser* program (Graf, et al., 1997).

Task Model: What tasks/items will facilitate data acquisition and evaluation? The underlying ECD framework should provide a solid basis for defining a task model or set of models. To extend its capabilities, this model should be standardized and modularly designed, supporting plug-and-play capabilities, and containing clean operational definitions of relevant knowledge and skill types. Examples and guidelines must be explicated and online tools developed to support item and task generation. An additional benefit would be to support either author- or auto-generated items.

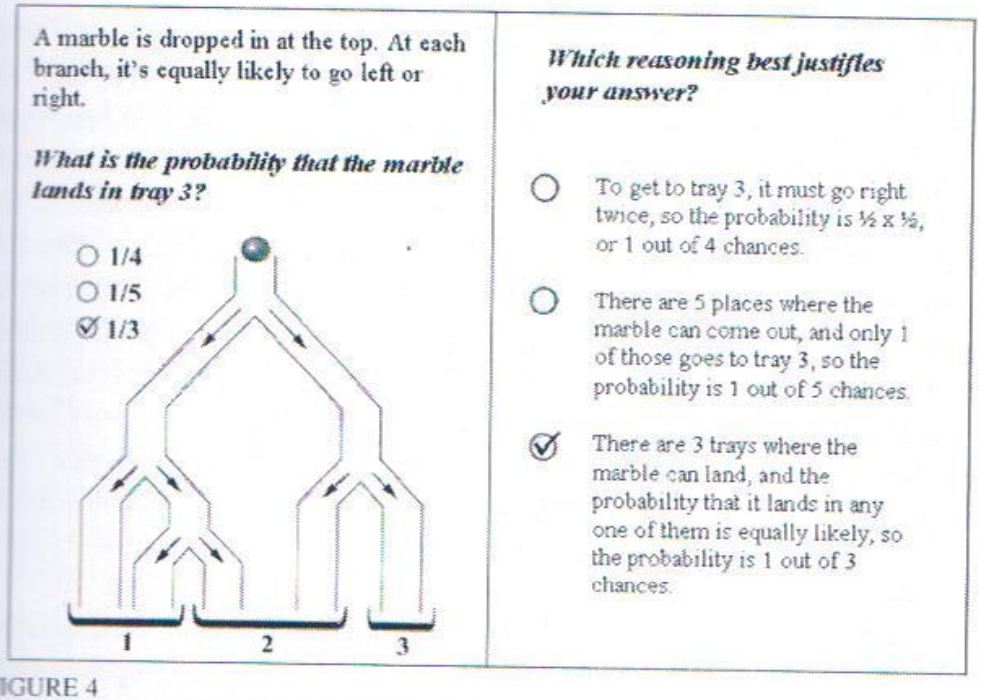

FIGURE 4
Diagnostic assessment in the area of probability.

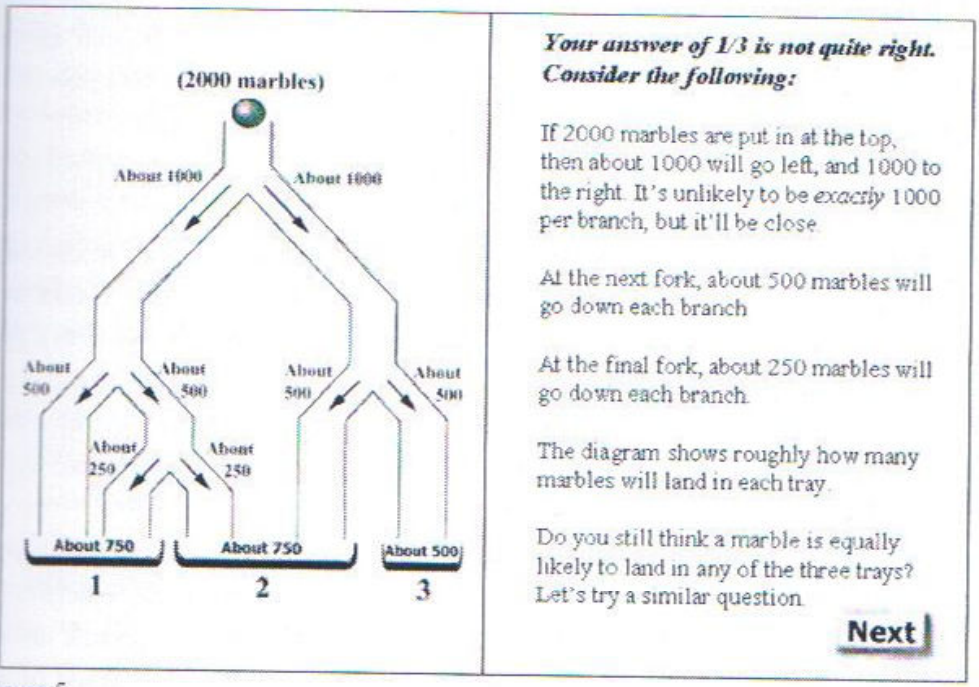

Figure 5
Diagnostic feedback in the area of probability.

To illustrate with the "sequences" content shown in Figure 2, here are some of its task specifications. First, the *presentation specifications* from the task model (for instructions, stimuli, stems, and options) include: text, numbers, symbols, tables, and relevant visual patterns. *Work product specifications* include both multiple choice and constructed response types. Finally, the *task model variables* include a number of attributes and associated values/ranges. Some of these are: (a) nature of the sequence (i.e., arithmetic, geometric, and other recursive), (b) nature of the task (e.g., find a term from a rule, induce a rule), (c) stimuli and response formats (e.g., list of numbers, symbols, or visual patterns; mathematical expression; verbal description; chart or table), (d) complexity of the underlying rule (simple to complex), (e) nature of numbers in the sequence (integers, fractions, decimals), (f) number of terms in the sequence (0 to >5), and so forth. This serves to sketch out the scope and nature of tasks to be created for the diagnostic assessment(s).

The *purpose* for which the assessment is intended may additionally influence task creation and presentation. For instance, the teacher may simply be concerned about the students' acquisition of a particular skill that was recently instructed. Alternatively, she may want to assess students' conceptual understanding on a given topic, or cover a wide range of knowledge types for some portion of the curriculum. The nature and format of the assessment should align with its purpose. Following is a brief discussion on knowledge types.

Flavors of Proficiencies (Three Types of Knowledge)

While we recognize that there are many different ways to classify knowledge (Anderson, 1983; Merrill, 1994; 2000; Shute, 1994; 1995), one reasonable approach is to classify knowledge into three types—basic knowledge (BK), procedural knowledge/skill (PK), and conceptual knowledge (CK). These three knowledge types strike an efficient balance between richness and utility in the task model. Again, the goal of an assessment may be to focus on three types equally; or on just one or two knowledge types.

Each knowledge type has differential, optimal instructional and assessment requirements. Given the aforementioned division over whether to focus more on "back to the basics" versus "improved conceptual understanding," it seems most reasonable to acknowledge the importance of the *range* of knowledge and skill types, each with its own optimal instruction and assessment format.

Basic Knowledge (BK)—What. Basic knowledge includes definitions, symbols, icons, formulas, or events. This type of knowledge answers the

question "WHAT?"—What does this symbol or icon represent? What is the definition of <some term>? What is the capital of Peru? What is the formula for calculating the arithmetic Mean? What is the atomic symbol for carbon in the periodic table, and what is its atomic number? BK requires an individual to know about, and discriminate among different things based on certain defined characteristics, thus BK assessment implies that the learner should be able to recognize, classify, sort, or produce some formula, basic definition, rule, and so on.

Procedural Knowledge (PK)—How. Procedural knowledge refers to the specific steps or actions needed in order to achieve a certain goal or perform a certain task. It is the representation and/or delineation of an operation or process and includes the conditions that apply or the decision rules related to performing steps within a procedure. PK answers the question "HOW?"—How do I achieve my goal? What are the steps, or what is the process that will help me achieve my goal? PK assessment requires that the learner should be able to actually accomplish some procedure or apply a rule, not simply recognize those things.

Conceptual Knowledge (CK)—Why. Conceptual knowledge refers to an understanding of an abstract idea or organized sets of ideas and the rules that relate them. This knowledge, for example, may be of a system or process—how that process works and the effects of change on or within a system. Or, this knowledge may be of a principle or strategy—when and why to employ a technique. CK answers the question "WHY?"—not simply a decision rule for performing a specific step, rather an understanding of the fundamental rationale for making that decision. CK assessment typically requires the learner to transfer BK and PK to novel areas, explaining some system or phenomenon, predicting some outcome, or strategizing a solution.

Putting it all together

The working premise is that *assessment results can and should have important implications for instruction*—positively influencing both the teaching and learning sides of the story. In today's classrooms, however, assessments are too often used for purposes of grading, promotion, and placement, but not for learning. The stance taken on assessments in this paper is that they should: (a) support—not undermine—the learning process for students and teachers, (b) provide more formative—compared to summative—information, i.e., useful feedback during the learning process rather than a single judgment at the end, and (c) be responsive to what we know about how children learn. In line with this, we are designing tools to automate and hence expedite the ECD approach

for assessment design that will support low- and high-stakes testing applications, as well as the derivation and presentation of valuable diagnostic information from such assessments. This is intended for students, teachers, parents, administrators, and others who need to monitor, foster, and report student growth from year to year.

The diagnostic assessments may be used alone, or linked to relevant modular, adaptable, and adaptive[1] instructional objects. Furthermore, teachers should be able to easily modify these objects (e.g., add additional problems or change content). The assessment and instructional objects will be based on task and instructional models, respectively, which together comprise a general content model. Reports generated at the end of a diagnostic assessment will thus be able to clearly specify students' strong and weak knowledge and skills, as well as prescribe specific instructional content based on the diagnostic report.

Automating the acquisition of proficiencies, claims, and evidences

The ideas described in this paper may serve as a first step in a multi-step path towards automating an assessment design process in line with an ECD approach. This can eventually inform the design of a learning (or test) management system. An existing knowledge-elicitation tool (i.e., Decompose, Network, Assess, aka DNA, as described in Shute, Torreano, & Willis, 2000; Shute & Torreano, 2003) already significantly reduces the laborious process of determining knowledge structures that map onto the claims of the proficiency model—the first part of the ECD process.

One of the salient features of DNA is that it is broadly applicable across domains. That is, it is equally capable of analyzing task performance (e.g., interpreting radar signals), as well as domains more conceptual in nature (e.g., understanding the factors that influence stock market fluctuations). As such, DNA represents a simple program for eliciting and organizing knowledge and skills from subject-matter experts. The result from the elicitation is a collection (database) of structured curriculum elements (i.e., learning objectives or claims) that comprise the basis for assessment, cognitive diagnosis, and instruction or feedback, as needed.

In general, DNA is intended to Decompose a domain into constituent knowledge and skill elements/objects, Network those elements into comprehensive structures, and employ other experts in the given domain to Assess

[1]The difference between "adaptable" and "adaptive" refers to applications that can be configured by a user (e.g., teacher who wants to edit the content), or whether it's coded to adjust itself to suit particular characteristics of the learner, respectively.

the validity, completeness, and reliability of the knowledge representation(s). In short—its main goal is to obtain the basic information for populating the proficiency model.

The program uses a semi-structured series of questions aimed at extracting and organizing knowledge structures from experts, in either a depth- or breadth-first manner. The questions correspond to the three main types of knowledge described above, namely: basic (AKA declarative or symbolic), procedural, and conceptual. Moreover, each knowledge type has its own path or interface in DNA. For example, eliciting basic knowledge invokes an interface that captures definitions, associated multi-media files, and other relevant information, while the procedural interface is more rule-based, such that the user can delineate steps, procedures, sub-procedures, conditionals, and so forth.

The first modification to DNA began with a project that served to align that program's underlying process model (see Figure 6, below) with that of ECD. Additional expansions on the DNA program would be to require it to elicit associated evidences, per knowledge and skill element that would be associated with particular nodes within the proficiency model. Incorporating a content model into the ECD framework (i.e., broadening the task model to accommodate instruction as well) will ensure that the associated assessments and/or instructional content are directly linked, via an evidentiary chain, to the desired proficiencies. Finally, conforming to current industry standards (e.g., IMS, SCORM) will allow these diagnostic assessments and instructional units to be used again in many different learning environments.

Developing Diagnostic Tasks

A set of claims within the proficiency model provides the benchmark by which to judge student learning and provide diagnostic information. But to move beyond "just a score," we need to know *why* a learner responded as he or she did. How do we obtain these claims in the first place? Again, one idea is to use an expanded version of DNA to elicit the knowledge and skill elements that will make up claims corresponding to the curriculum or course. These claims of student proficiencies represent what is expected of a student at the end of the course (or instructional episode). To be most useful for diagnostic purposes, the claims should be arrayed in a hierarchical network, if possible. This will allow the computer to work backwards from a given problem to unearth the root cause. Next, to obtain relevant "evidences" of students' mastery (or non-mastery) of particular claims, we can specify a range of work products (e.g., multiple-choice answer key, constructed or patterns of responses, etc.). Finally, we need to delin-

eate particular learning indicators with salient features, per claim (e.g., If a student does X, then she knows Y, and the probability is p = .Z that she is "high" on this proficiency). This would involve setting explicit links between evidence(s) and proficiency model variables.

Task development can ensue from task models, which in turn ensue from the specification of evidences that are needed. The task models can assist in the production of valid assessments in which learners will act, and work products will result. The focus here is on specifying the evidentiary value of tasks, and the support for test assembly specifications. These evidences will mediate the relationship between tasks and proficiency model variables. Finally, reporting rules will need to be developed in order to summarize inferences of students' proficiencies. These reports can be shared with users, and should be written in a clear and informative style. This will entail deriving a set of rules to collect information and summarize it succinctly. Additional rules are needed to aggregate data – i.e., to collapse up (to more general), or unpack down (to more specific), information, depending on the desired grain size of interest. This would include provisions to either generate reports or select remediation or new instructional modules that are relevant. This brings us to the question of: What are the building blocks that should be used to design and develop these valid diagnostic assessments?

Learning Objects—The building blocks

In order to support an assessment design approach and maintain educational use and re-use of the assessments, content may assume the form of learning objects (LOs). The size of a learning object can be defined with respect to the content's role in the assessment process. That is, each LO in the task model is the smallest chunk of material that exists as a coherent whole, independent of other learning objects, and still serves some definable role in the assessment process. This may be an HTML page, simulation, question with several multiple-choice answer options, reading a passage and responding to questions about it, and so forth.

The presentation of learning objects must not presume any predetermined sequencing of the objects, thus must be modular (i.e., self-contained) with internal tags reflecting information about the object's format, purpose, size, knowledge type, related LOs, and other attributes. Furthermore, each claim in the proficiency model may have an associated *family of learning objects*. Each family member will collectively, yet uniquely, relate to the learning objective (claim). Our task model can provide requirements for the learning object family, such as the number and type of objects contained within it, difficulty indices, and so forth.

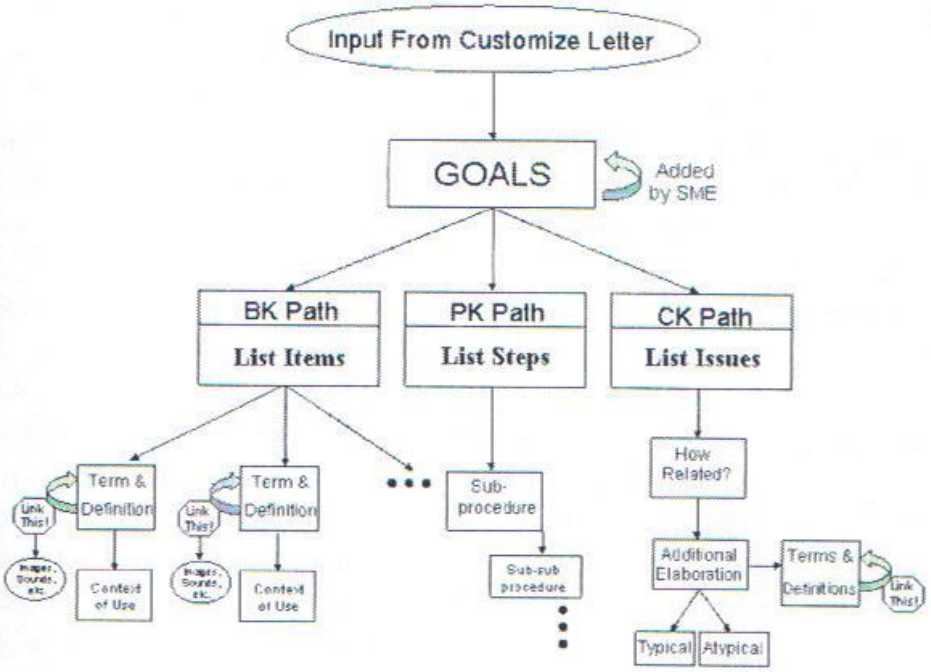

Figure 6
Object model representing DNA's Decompose Module (from Shute & Torreano, 2003).

In short, claims are part of the proficiency model, and LOs are the online manifestations of claims. Thus, claims are assessed in terms of inferences based on learners' performance in relation to LOs. Each claim may be instructed and assessed via different types of objects such as instructional objects (IOs) and assessment objects (AOs).

To illustrate this in the domain of probability, suppose the claim was that the student will "know about permutations." One AO could assess the learner's ability to recall the definition of permutation; a related AO could require the student to list all pair-wise permutations of a set of three numbers. The resulting diagnosis—the estimate of mastery, per claim—determines the tutorial action, or curricular flow, for a particular learner. For example, if a claim has achieved mastery status, the system would present/assess content from a new section of the curriculum. If mastery was not attained, the system would either (a) present remedial instruction in relation to the current claim, or (b) require the student to continue to solve problems involving the current claim, perhaps in the context of easier AOs. The teacher or instructional designer sets mastery thresholds in advance, and executive control rules can be used to determine whether remediation or continued problem solving is indicated.

SUMMARY

In summary, the U.S. educational system has some problems. These will be exacerbated if we persist in using assessments in a summative (single score) manner rather than in formative ways. Researchers at ETS have been working on an assessment design framework called ECD. This assists in creating valid assessments and may be used to create diagnostic assessments. Furthermore, automating this process is possible. This paper has described one tool that may be modified and used to elicit information that feeds directly into the student and evidence models. The first capability (eliciting proficiencies) already exists, but the second category (eliciting evidences) does not. One idea is to attach this elicitation to the claim-extraction probes (i.e., for each successive claim elicited, one could simply add follow-on queries relating to evidences for that knowledge or skill). The goal would be to seamlessly link this capability to other tools that support the ECD framework, such as scoring tools from ECD, relevant task models, and item-authoring shells and models. We would then have the pieces in place for expediting the development of a formative, diagnostic assessment solution. This research defines the juxtaposition of basic research with applied considerations of use.

The future of testing will undoubtedly embrace assessments that concurrently, or subsequently, support learning. The model-driven, diagnostic assessment system defined herein can provide clear examples of emerging competence, highlight gaps in understanding, and suggest activities for improved learning. Teachers and trainers who use the results of these assessments will be able to assist individual students more readily and precisely. And districts that consider assessment information of this sort, in conjunction with district and state tests, position themselves to make informed decisions about program planning, resource allocation, and teacher professional development. The results of such research efforts will allow us to respond to the current and emerging educational needs mentioned earlier, that could inform and enhance teaching and learning.

The basic idea is to explicitly connect instructional activities and assessments to an underlying proficiency model for each content area. This allows one to pinpoint an individual student's current position on a continuous, empirically determined proficiency scale. It also can provide detailed and accurate information on student performance for teachers, students, administrators, and parents. This information can then be used to select and provide instructional interventions that are specifically targeted at the needs of

a particular student or group of students. Teachers will thus be able to monitor student progress regularly using a variety of assessment methodologies, many of which will be unobtrusive and fully integrated into the instructional milieu. This will provide the teacher with immediate feedback on what is working in the classroom and what is not, as well as which students are making progress, and which ones are not. This timely feedback will also provide the educator with the unprecedented ability to intercede early in the learning cycle and adjust the instructional program as needed to meet the various requirements of students. By providing frequent and accurate feedback, teachers can keep students moving forward and avoid leaving any child behind.

In conclusion, school communities must begin to use assessment results in a formative way to determine how well they are meeting instructional goals and how to alter curricula and instruction so that goals can be better met. As Landgraf (2001) pointed out, "Well-designed tests tied to standards and curriculum can provide useful information to guide instruction and help students learn. Test results can also provide useful data to guide sound education policy decisions" (p. 3). This paper represents an attempt to accomplish those goals by automating the formulation of diagnostic assessments and eventually companion instructional units, in the service of learning.

Acknowledgements

I'd like to thank Aurora Graf, Gail Baxter, Jody Underwood, Dan Eignor, and Malcolm Bauer for contributing to the ideas expressed in this paper. An earlier version of this paper was presented at a workshop for the Intelligent Tutoring Systems (ITS) conference in 2002 (Biarritz, France).

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, 9(5), ISSN: 1068-2341.

Graf, E. A. (2003, September). *Designing a Proficiency Model and Associated Item Models for an 8th Grade Mathematics Unit on Sequences*. Paper presented at the Math Forum (Sept. 25, 2003), ETS, Princeton, NJ.

Graf, E. A., Greer, B., Neal, D., Durbin, Y., Lenges, A., Minstrell, J., Hunt, E., Levidow, B., Aiken, D., & Traynor, C. (1997). *Proportional Reasoning Module of DIAGNOSER 6.0* [Computer Program]. Hunt Lab. Retrieved, 2002, from the World Wide Web: http://depts.washington.edu/huntlab/diagnoser/download/index.html/.

International Study Center at Boston College. (1998). *Third International Mathematics and Science Study (TIMSS) Highlights from the Final Year of Secondary School*. Available from: http://timss.bc.edu/timss1995i/HiLightC.html

Landgraf, K. (March 2001). Using Assessments and Accountability to Raise Student Achievement. Testimony before the Education Reform Subcommittee of the House Committee on Education and the Workforce on Measuring Success.

Merrill, M. D. (1994). *Instructional Design Theory*. Englewood Cliffs: Educational Technology Publications.

Merrill, M. D. (2000). Knowledge objects and mental models. In D. A. Wiley (Ed.), *The Instructional Use of Learning Objects: Online Version*. Retrieved March 6, 2002, from the World Wide Web: http://reusability.org/read/chapters/merrill.doc.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13-23.

Minstrell, J. (2000). Student thinking and related instruction: Creating a facet-based learning environment. In J. Pellegrino, L. Jones, & K. Mitchell (Eds.) *Grading the Nation's Report Card: Research for the Evaluation of NAEP*. Committee on the Evaluation of NAEP, Board on Testing and Assessment. Washington, DC: National Academy Press.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2000). Bayes Nets in Educational Assessment: Where Do the Numbers Come From? *CSE Technical Report 518*.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Discussion*

Mislevy, R.J., Steinberg, L.S., Almond, R.G., Haertel, G., & Penuel, W. (in press). Leverage points for improving educational assessment. In B. Means & G. Haertel (Eds.), *Evaluating the effects of technology in education.*

National Center for Educational Statistics. (2001a). *Highlights from the Third International Mathematics and Science Study-Repeat (TIMSS-R)*. U.S. Department of Education. Retrieved, from the World Wide Web: http://nces.ed.gov/timss/timss-r/highlights.asp#1.

Office of Educational Research and Improvement. (2001). Statistical Highlights of Student Achievement (Memo): U.S. Department of Education.

Shute, V. J. (1994). Learning processes and learning outcomes. In T. Husen & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education* (2nd Edition) (pp. 3315-3325). New York, NY: Pergamon Press.

Shute, V. J. (1995). SMART: Student Modeling Approach for Responsive Tutoring. *User Modeling and User-Adapted Interaction*, 5, 1-44.

Shute, V. J. & Torreano, L., & Willis, R. (2000). DNA: Towards an automated knowledge elicitation and organization tool. In S. P. Lajoie (Ed.) *Computers as Cognitive Tools, Volume 2*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 309-335.

Shute, V. J., & Torreano, L. A. (2003). Formative Evaluation of An Automated Knowledge Elicitation and Organization Tool. In T. Murray, S. Ainsworth, & S. Blessing (Eds.), *Authoring Tools for Advanced Technology Learning Environments: Toward Cost-Effective Adaptive, Interactive, and Intelligent Educational Software* (pp. 149-180). Kluwer Academic Publishers. Printed in the Netherlands.

Snow, C.E. & Jones, J. (2001). Making a silk purse. *Education Week Commentary*, April 25, 2001.