RESEARCH REPORT

July 2003
RR-03-16

A Brief Introduction to Evidence-centered Design

Robert J. Mislevy

Russell G. Almond

Janice F. Lukas

Research &
Development Division
Princeton, NJ 08541

A Brief Introduction to Evidence-centered Design

Robert J. Mislevy, University of Maryland

Russell G. Almond and Janice F. Lukas, Educational Testing Service, Princeton, NJ

July 2003

Abstract

Evidence-centered assessment design (ECD) is an approach to constructing educational assessments in terms of evidentiary arguments. This paper provides an introduction to the basic ideas of ECD, as well as some of the terminology and models that have been developed to implement the approach. In particular, it presents the high-level models of the Conceptual Assessment Framework (CAF) and the Four-process Delivery Architecture for assessment delivery systems. Special attention is given to the role of probability-based reasoning in accumulating evidence across task performances, in terms of belief about unobservable variables that characterize the knowledge, skills, and/or abilities of students. This is the role traditionally associated with psychometric models, such as those of item response theory (IRT) and latent class models. To unify the ideas and to provide a foundation for extending probability-based reasoning in assessment applications more broadly, however, a more general expression in terms of graphical models is indicated. This brief overview of evidence-centered design provides the reader with a feel for where and how graphical models fit into the larger enterprise of educational and psychological assessment. A simple example based on familiar large-scale standardized tests such as the Graduate Record Examinations® (GRE®) is used to fix ideas.

Key words: Assessment design, delivery system, evidence, psychometrics

Table of Contents

List of Figures

Page

Overview

What all educational assessments have in common is the desire to reason from particular things students say, do, or make, to broader inferences about their knowledge and abilities. Over the past century, a number of assessment methods have evolved for addressing this problem in a principled and systematic manner. The measurement models of classical test theory and, more recently, item response theory (IRT) and latent class analysis, have proved quite satisfactory for the large scale tests and classroom quizzes with which every reader is by now quite familiar.

But off-the-shelf assessments and standardized tests are increasingly unsatisfactory for guiding learning and evaluating students' progress. Advances in cognitive and instructional sciences stretch our expectations about the kinds of knowledge and skills we want to develop in students and the kinds of observations we need to evidence them (Glaser, Lesgold, & Lajoie, 1987). Advances in technology make it possible to evoke evidence of knowledge more broadly conceived and to capture more complex performances. One of the most serious bottlenecks we face, however, is making sense of complex data that result.

Fortunately, advances in evidentiary reasoning (Schum, 1994) and in statistical modeling (Gelman, Carlin, Stern, & Rubin, 1995) allow us to bring probability-based reasoning to bear on the problems of modeling and uncertainty that arise naturally in all assessments. These advances extend the principles upon which familiar test theory is grounded to more varied and complex inferences from more complex data (Mislevy, 1994). One cannot simply construct "good tasks" in isolation, however, and hope that someone down the line will figure out "how to score them." One must design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them (Messick, 1994). More complex statistical models may indeed be required, but they evolve from the substance of the assessment problem, jointly with the purposes of the assessment and the design of the tasks.

The evidence-centered design (ECD) project at Educational Testing Service (ETS) provides a conceptual design framework for the elements of a coherent assessment, at a level of generality that supports a broad range of assessment types, from familiar standardized tests and classroom quizzes, to coached practice systems and simulation-based assessments, to portfolios and student-tutor interaction. The design framework is based on the principles of evidentiary reasoning (Mislevy, Steinberg, & Almond, 2003) and the exigencies of assessment production

and delivery (Almond, Steinberg, & Mislevy, 2002). Designing assessment products in such a framework ensures that the way in which evidence is gathered and interpreted is consistent with the underlying knowledge and purposes the assessment is intended to address. The common design architecture further ensures coordination among the work of different specialists, such as statisticians, task authors, delivery-process developers, and interface designers. While the primary focus of measurement specialists is building, fitting, testing, and reasoning from statistical models, this primer places such models into the context of the assessment process. It will serve to motivate, we hope, and to lay the groundwork for more technical discussions of both the framework and its applications.

In accordance with the goal, most attention is focused on models called the Conceptual Assessment Framework, or CAF, and the Four-process Delivery Architecture for assessment delivery systems. The reader interested in a fuller treatment of ECD is referred to Mislevy et al. (2002) for connections to the philosophy of argument and discussions of the earlier stages of design, and to Almond et al. (2002) for amplification on delivery system architecture.

The first section provides a rationale for assessment as a special case of an exercise in evidentiary reasoning, with validity as the grounds for the inferences drawn from assessment data (Cronbach, 1989; Embretson, 1983; Kane, 1992; and Messick 1989, 1994). ECD provides a structural framework for parsing and developing assessments from this perspective. A brief overview of the CAF and Four-process Delivery Architecture are presented.

As running illustrations, we will use examples based on the paper and pencil (P&P) and the computer adaptive (CAT) versions of the Graduate Record Examinations® (GRE®), and from a prototype of a computer-based assessment of proficiency in dental hygiene (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999, 2002), developed by ETS and the Chauncey Group International for the Dental Interactive Simulations Corporation (DISC). As this is written, the GRE comprises three domains of items, concerning Verbal, Quantitative, and Analytic reasoning skills. In each case, a student responds to a number of items in the domain, and an estimate of a single proficiency with regard to that domain is reported. The DISC prototype concerns seven aspects of knowledge and skill, and evidence is gathered as the student works through scenarios based on the examination and treatment of simulation patients.

Assessment as Evidentiary Argument

Advances in cognitive psychology deepen our understanding of how students gain and use knowledge. Advances in technology make it possible to capture more complex performances in assessment settings by including, for example, simulation, interactivity, collaboration, and constructed response. Automated methods have become available for parsing complex work products and identifying their educationally meaningful features. The challenge is in knowing just how to put all this new knowledge to work so that it best serves the assessment's purposes. Familiar schemas for designing and analyzing tests produce assessments that are useful because they are coherent, but they are limited by the constraints under which they evolved. Breaking beyond the constraints requires not only the means for doing so (through advances such as those mentioned above), but schemas for producing assessments that are again coherent; that is, assessments that may indeed gather complex data to ground inferences about complex student models and to gauge complex learning or evaluate complex programs—but that build on a sound chain of reasoning from observation to inference.

Recent work on validity in assessment lays the conceptual groundwork for such a scheme. The contemporary view focuses on the support—conceptual, substantive, and statistical—that assessment data provide for inferences or actions (Messick, 1989). From this view, an assessment is a special case of evidentiary reasoning. Messick (1994) lays out its general form in the following quotation:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

This perspective is valuable because it helps organize thinking about assessments for all kinds of purposes, using all kinds of data, task types, scoring methods, and statistical models. We can ask of a simulation task, for example, just what knowledge and skills is it meant to reveal?

Do the scoring methods pick up the clues that are present in performances? How is this evidence synthesized across multiple tasks, or compared when different students attempt different tasks? Every decision in the assessment design process influences the chain of reasoning from examinees' behaviors in the task setting to conclusions about their knowledge or abilities. From an evidentiary reasoning perspective, we can examine the impact of these decisions on the inferences we ultimately need to draw.

As powerful as it is in organizing thinking, simply having this conceptual point of view isn't as helpful as it could be in carrying out the actual work of designing and implementing assessments. A more structured framework is needed to provide common terminology and design objects that make the assessment design explicit and link the design elements to the processes inherent in an operational assessment. Such a framework not only makes the underlying evidentiary structure of an assessment more explicit, but also makes the operational elements of an assessment easier to reuse and to share. The ECD design models address this need.

Basic ECD Structures

While the full ECD framework starts with the initial analysis of a substantive domain and assessment purposes, we will focus in this overview on the two parts that are closest to the implemented assessment, the CAF, and the Four-process Delivery Architecture for assessment delivery systems. Again, the interested reader is referred to Mislevy, Steinberg, and Almond (2002) for a discussion of the design process from start to finish. Suffice it to say that in any particular assessment the objects in the CAF models described in general terms here will need to have been designed to address the purposes of that particular assessment. In line with the Messick quotation cited earlier, all of the task characteristics were selected to provide the opportunity to elicit evidence about the targeted knowledge and skill; all of the scoring procedures were designed to capture, in terms of observable variables, the features of student work that are relevant as evidence to that end; and the characteristics of students that are reflected as student model variables were meant to summarize evidence about the relevant knowledge and skills from a perspective and at a grain size that suits the purpose of the assessment.

Figures 1 and 2 depict the main ECD models for design and delivery respectively, that is, the CAF and the Four-process Delivery Architecture. We will walk through their contents and their connections in the following sections. In a nutshell, the CAF models lay out the blueprint for the operational elements of an assessment, and their interrelationships coordinate its substantive, statistical, and operational aspects. The CAF models provide the technical detail required for implementation: specifications, operational requirements, statistical models, details of rubrics, and so on. The four processes of the delivery system carry out, examinee by examinee, the functions of selecting and administering tasks, interacting as required with the examinee to present materials and capture work products, and then evaluating responses from each task and accumulating evidence across them.

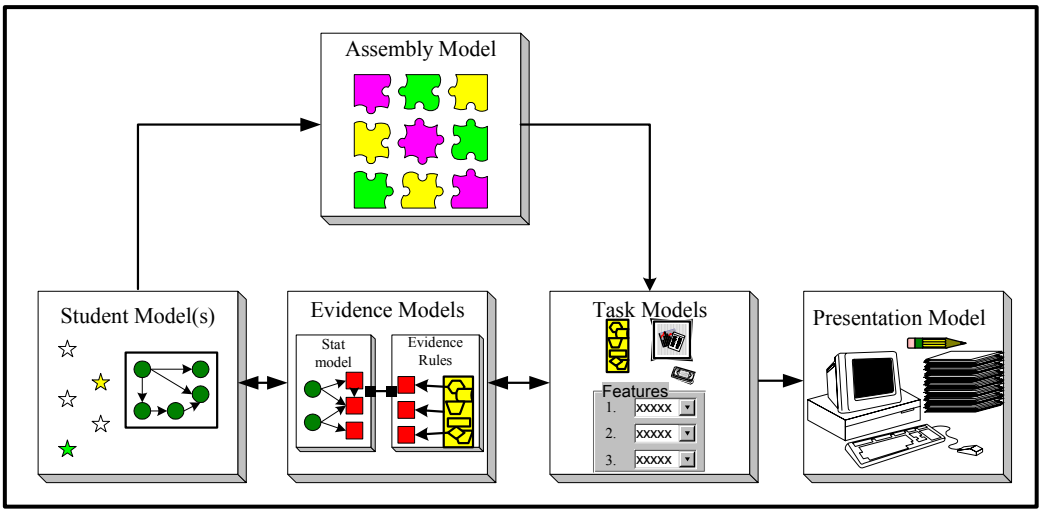

Figure 1. **The principal design objects of the Conceptual Assessment Framework (CAF).** These models are a bridge between the assessment argument and the operational activities of an assessment system. Looking to the assessment argument, they provide a formal framework for specifying the knowledge and skills to be measured, the conditions under which observations will be made, and the nature of the evidence that will be gathered to support the intended inference. Looking to the operational assessment, they describe the requirements for the processes in the assessment delivery system.

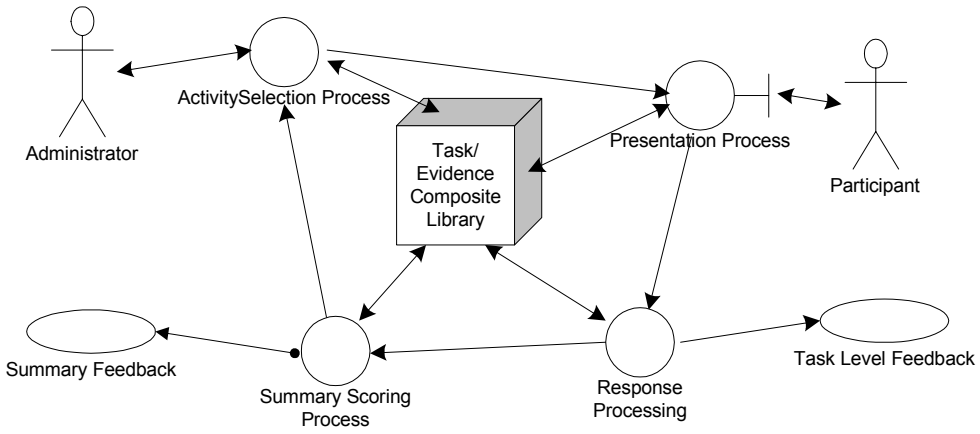

Figure 2. **The four principal processes in the assessment cycle.**
The Activity Selection Process selects a task (tasks include items, sets of items, or other activities) and directs the Presentation Process to display it. When the participant has finished interacting with the item, the Presentation Process sends the results (a Work Product) to Response Processing. This process identifies essential Observations about the results and passes them to the Summary Scoring Process, which updates the Scoring Record, tracking our beliefs about the participant's knowledge. All four processes add information to the Results Database. The Activity Selection Process then makes a decision about what to do next, based on the current beliefs about the participant or other criteria.

The Conceptual Assessment Framework

The blueprint for an assessment is called the *Conceptual Assessment Framework*. To make it easier to rearrange the pieces of the framework (and deal with one at a time when appropriate), the framework is divided into a number of pieces called *models*. Each model provides specifications that answer such critical questions as "What are we measuring?" or "How do we measure it?"

What Are We Measuring: The Student Model

A *Student Model* defines one or more variables related to the knowledge, skills, and abilities we wish to measure. A simple student model characterizes a student in terms of the proportion of a domain of tasks the student is likely to answer correctly. A more complicated model might characterize a student in terms of degree or nature of knowledge of several kinds, each of which may be required in different combinations in different tasks. Looking ahead, the

student model variables will be a subset of the variables in a graphical model that accumulates evidence across tasks.

In each of the three GRE domains, such as Verbal Reasoning, the student model consists of a single unobservable variable, a proficiency in that domain. Any student's value on this variable is not known and indeed can never be known with certainty. At any point in time, our state of knowledge about its value is expressed by a probability distribution across the range of values it might take. Figure 3 depicts the student model for this example: a single proficiency variable, denoted θ, and a probability distribution, represented by a table meant to suggest a probability distribution. In the P&P GRE, a student is administered a preassembled test containing over 100 test items. She answers them all, and her θ values for the three areas are estimated based on her responses to the items in each of the three skill domains. In the computer adaptive test (CAT) form of GRE, items are selected one at a time to be presented, in each case based on the student's previous responses in order to be more informative about the value of her θ.

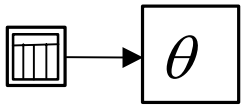

Figure 3. **The Student Model for a GRE measure.**
The student model for a GRE measure (e.g., Verbal, Quantitative, or Analytic) is a single variable that characterizes a student's proficiency, denoted θ, in that domain; that is, her tendency to provide correct answers to items in that measure. The small box to the left of θ represents a probability distribution that expresses what is known about a student's inherently unobservable θ at a given point in time.

At the beginning of an examinee's assessment, the probability distribution representing her status will be uninformative. We will update it in accordance with how the examinee performs in various situations we have structured; that is, when we see her responses to some GRE Verbal test items. Later, we will look at student models with several variables, each representing some aspect of knowledge, skill, or ability posited to influence students' performance. In each case, however, the idea is the same as in the simple IRT case: These variables are how we characterize students' knowledge. We don't get to observe them directly,

so we express what we do know about them in terms of a probability distribution, and evidence in the form of behavior in assessment situations allows us to update our knowledge, by updating the probability distributions accordingly.

Figure 4 shows a more complex student model for the DISC prototype assessment, one that has variables for six areas of knowledge in the domain of dental hygiene, along with an overall proficiency. Testing scenarios based on interactions with simulated patients would provide information about two to four of these variables.

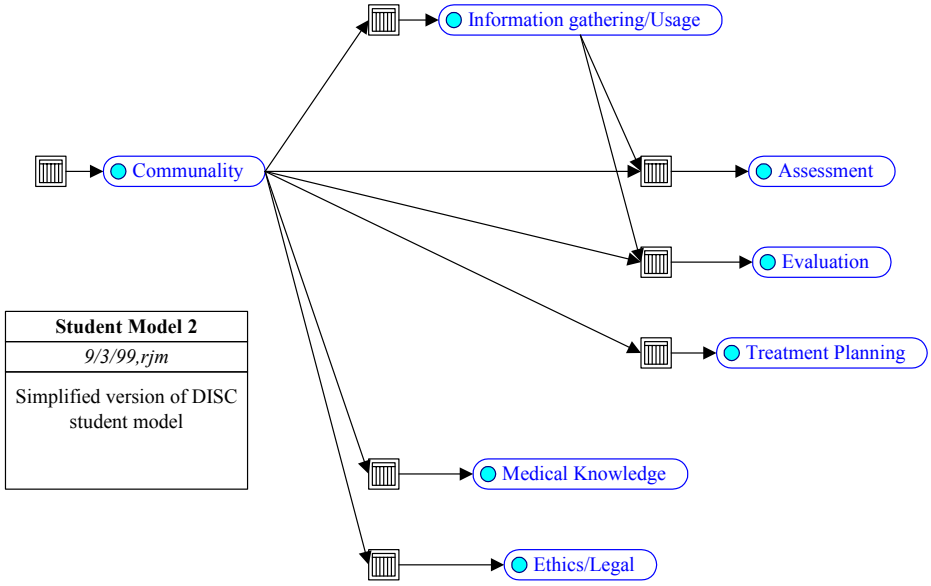

Figure 4. **The Student Model for a simulation-based assessment of problem-solving in dental hygiene.**
The student model variables are used to synthesize evidence from task performance, in terms of a probability distribution over them. Arrows represent associations among the aspects of knowledge and skill.

How Do We Measure It: The Evidence Model

Evidence Models provide detailed instructions on how we should update our information about the student model variables given a performance in the form of examinees' *work products* from tasks. An evidence model contains two parts, which play distinct roles in the assessment argument:

- *Evidence Rules* describe how *observable variables* summarize an examinee's performance in a particular task from the *work product* that the examinee produced for

that task. These observables are the primary outcomes from tasks, and they provide both information that will be used to update our beliefs about student model variables and information that will be used for task level feedback. In an operational assessment, evidence rules guide the *Response Scoring* process. It is important to note that evidence rules concern the *identification and summary of evidence within tasks*, in terms of observable variables.

For GRE items, the observable variable associated with each item is whether it is answered correctly or incorrectly. Its value is determined by comparing the student's response with the answer key: correct if they match, incorrect if they don't.

For DISC tasks, there are generally several observable variables evaluated from each complex task performance. In scenarios involving the initial assessment of a new patient, for example, there are five observables, including ratings of qualities of "Adapting to situational constraints" and "Adequacy of patient-history procedures." The evaluation rules are based on whether the examinee has carried out an assessment of the patient that addresses issues that are implied by the patient's responses, condition, and test results such as radiographs and probing depths.

The *Measurement Model* part of the evidence model provides information about the connection between *student model variables* and *observable variables*. Psychometric models are often used for this purpose, including the familiar classical test theory and IRT models, and the less familiar latent models and cognitive diagnosis models. In an operational assessment, measurement models guide the *Summary Scoring* process and concern the *accumulation and synthesis of evidence across tasks*, in terms of student model variables.

Looking ahead again, a graphical model containing both the student model variables and observable variables is the machinery that effects probability-based accumulation and synthesis of evidence over task performances. For our GRE example, the measurement model is IRT. Figure 5 shows the measurement model used in the GRE-CAT. It gives the probability for a correct or incorrect response to a particular Item j, as a function of a student's IRT proficiency variable, θ. When it comes time to update belief about a student's θ based on a response to this item, this fragment is joined with the student model discussed earlier, and the updating procedures discussed in Mislevy

(1994), for example, enter into play. Figure 6 depicts a measurement model for a more complex DISC task, in which five aspects of performance are captured as observable variables, and two aspects of proficiency are updated in terms of probability distributions for student-model variables.

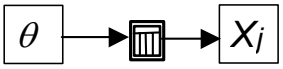

Figure 5. **The Measurement Model used in GRE-CAT.**
This figure shows that the probability distribution of the variable for the response to Item j, or X_j, depends on the student's proficiency variable θ. When a response to X_j is observed one uses Bayes Theorem to update belief about θ, in terms of its probability distribution in the student model.

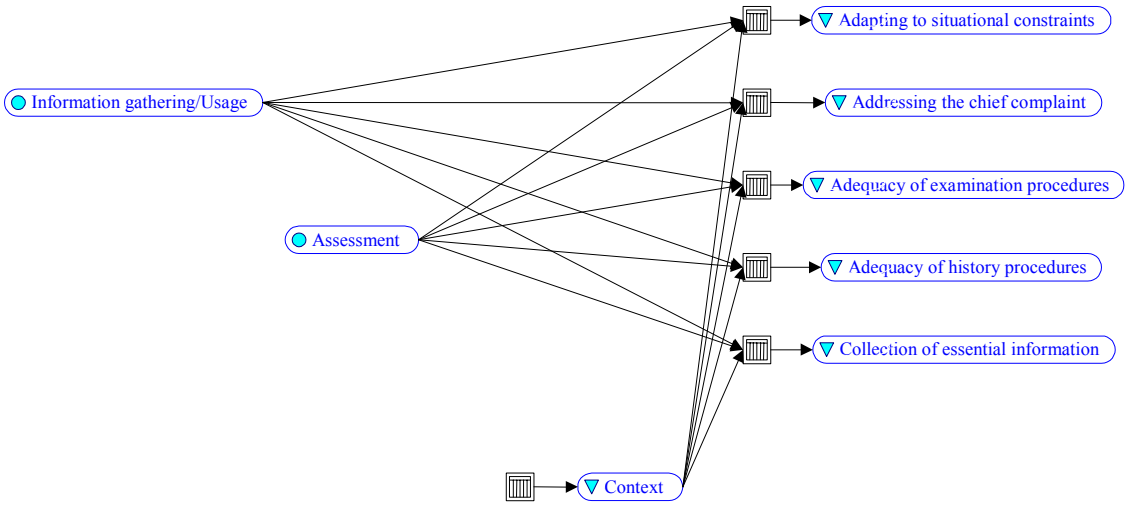

Figure 6. **The Measurement Model used in the DISC prototype.**
This is a measurement model for scenarios in which an examinee is examining a new patient, and there are no special medical or ethical considerations. The five variables on the right represent observable variables. The two toward the upper left are two of the variables from the student model, which are modeled as governing, in probability, students' responses in these situations. The variable at the bottom center accounts for the dependencies among the observables that arise from evaluating multiple aspects of the same complex performance.

Where Do We Measure It: The Task Model

Task Models describe how to structure the kinds of situations we need to obtain the kinds of evidence needed for the evidence models. They describe the *presentation material* that is presented to the examinee and the *work products* generated in response. They also contain *task*

model variables that describe features of tasks as well as how those features are related to the presentation material and work products. Those features can be used by task authors to help structure their work, by psychometricians to help reduce the number of pretest subjects needed, and by test assemblers to help ensure that a particular form of the assessment is balanced across key kinds of tasks.

A task model does not represent a single task, but rather a family of potential tasks waiting to be written. Tasks are made from task models by filling in the specification made by the task model, that is, finding or authoring presentation material and setting the values of the task model variables. A typical assessment will have many task models representing different families of tasks. A family of tasks produced from a task model is not necessarily meant to be psychometrically interchangeable. That is, the tasks are not item clones for producing alternate forms, but are a structure for understanding and controlling evidential variation, often systematically manipulating the evidential value (and the statistical parameters) of the item for the assessment in question.

A task model in the GRE describes a class of test items. There is some correspondence between task models and GRE item types, such as sentence completion, passage comprehension, and quantitative comparison. Different item types will generally require different task models because different sets of variables will be needed to describe their distinct kinds of stimulus materials and presentation formats and because different features may be important in modeling item parameters or controlling item selection. Different task models will be required for P&P and CAT versions of what is the same item from the perspective of content, because specifications for presenting and managing the item are wholly different in the two modes.

A task model for DISC is a skeleton of a scenario involving a simulated patient. A particular task is created by determining the values of variables that characterize key aspects of the patient's condition, providing responses to patient history probes and dental procedures, and generating stimulus materials that correspond to the patient's etiology and condition.

How Much Do We Need to Measure: The Assembly Model

Assembly Models describe how the student models, evidence models, and task models must work together to form the psychometric backbone of the assessment. Targets describe how

accurately each student model variable must be measured, and constraints describe how tasks must be balanced to properly reflect the breadth and diversity of the domain being assessed.

In both the P&P and CAT versions of the GRE, assembly rules govern the mix of item types, the content of reading passages, numbers of items that use geometric figures, abstract versus concrete concepts, sentence complexity, and many other task features. Statistical features of items, such as their difficulty, are also taken into account, but in different ways depending on the way they are presented. GRE P&P test forms are all constructed to match the same targeted distributions of item difficulty and overall accuracy. GRE-CAT test forms are custom matched to each individual student to increase information about that student in light of their unfolding sequence of responses. Students doing well tend to be administered harder items, while students doing poorly tend to be administered easier ones.

At present, the DISC prototype does not have an assembly model. In practice, assembly would be governed by task model variables that address the difficulty of tasks, their coverage of stages of interaction with patients, and patient types such as geriatric and pediatric cases.

How Does It Look: The Presentation Model

Today's assessments are often delivered through many different means. For example, test takers might use paper and pencil, a stand-alone computer or the Web, or a hand-held device. Assessments might also be read aloud over the phone or consist of portfolios assembled by the students. A *Presentation Model* describes how the tasks appear in various settings, providing a style sheet for organizing the material to be presented and captured. The same GRE items can be administered under either P&P or CAT formats, but the ways they are composited or rendered on a computer screen require wholly different sets of instructions—in one case directions to the printer, in the other case code to the computer that will display them. The presentation model for DISC, not implemented as this is written, would be more complex. It has to address the setup of the simulation of the patient, accept and implement the student's queries and prescriptions, present information about the simulated patient, and capture the student's responses to the filling in of patient evaluation forms.

Putting It All Together: The Delivery System Model

The *Delivery System Model* describes the collection of student, evidence, task, assembly, and presentation models necessary for the assessment and how they will work together. It also describes issues that cut across all of the other models, such as platform, security, and timing.

Breaking the assessment specification up into many smaller pieces allows it to be reassembled in different configurations for different purposes. For example, a diagnostic assessment requires a finer grain size student model than a selection/placement assessment. If we want to use the same tasks in both the diagnostic and selection assessment, we must use the same task models (written generally enough to address both purposes). However, we will want different evidence models, each one appropriate to the level of detail consistent with the purpose of the assessment.

Four-process Delivery Architecture for Assessment Delivery

Assessments are delivered in a variety of platforms. Paper-and-pencil tests are still the most widely used, oral exams have a long history, and the use of computer-based tests is growing rapidly. New ways to deliver tests are making an appearance as well: over the Web, over the phone, and with hand-held devices.

To assist in planning for all these diverse ways of delivering a test, ECD provides a generic framework: the Four-process Delivery Architecture. The Four-process Delivery Architecture (shown in Figure 2) is an ideal system since any realized assessment system must contain these four processes in some form or other. They are essential to making the observations and drawing the inferences that comprise an assessment argument. This is true whether some of the processes degenerate or are collapsed or degenerate in a given system, and regardless of whether they are carried out by humans, computers, or human-computer interactions.

How Is the Interaction With the Examinee Handled: The Presentation Process

The *Presentation Process* is responsible for presenting the task and all supporting presentation material to the student and for gathering up the work products. Examples include a display engine for computer-based testing, a simulator that can capture an activity trace, and a system for distributing test booklets and capturing and scanning the answer sheets. In the P&P GRE, the presentation process concerns administering preassembled test booklets to examinees

and collecting their bubbled-in answer sheets. In the CAT, presentation involves presenting a customized sequence of items to an examinee one at a time, and after each item capturing the response so it can be evaluated on the spot to direct the selection of the next item. In an operational DISC assessment, presentation could be based on predetermined suites of cases or cases selected at random from a pool, subject to coverage constraints.

How Is Evidence Extracted From a Task Performance: Response Processing

Response Processing is responsible for identifying the key features of the work product or the observable outcomes for one particular task. The observable outcomes can either revert back to the participant for task-level feedback or on to the summary scoring process. Examples include matching a selected response to an answer key, running an essay through an automated scoring engine, or having a human rater score a student portfolio according to a rubric. The CAF evidence rules specify how this is to be accomplished. Response processing can consist of multiple stages, as when lexical and syntactic features are identified in an essay and a regression model is used to summarize them into a single score for a response to the task.

The logical content of response processing is the same in CAT and P&P GRE forms: The student's answer is compared with the key. The implementation is rather different for the two modes of administration, however. In the P&P version, sophisticated algorithms and sensitive machines are employed to determine, via relative intensities of light reflection, which answer bubble the student marked. In the CAT version, the character that corresponds to the location where the student clicked a mouse button to indicate an answer choice is compared with the character stored in memory as the key.

Response processing in an operational DISC assessment would consist of running automated rules on the sequence of actions a student carried out in order to identify and summarize salient features, as required, for example, to determine whether the student had addressed the patient's chief complaint. The students' filled-in evaluation forms for their simulation patients would also be evaluated for not only correctness, but efficiency and internal consistency as well.

How Is Evidence Accumulated Across Tasks: Summary Scoring

The *Summary Scoring* Process is responsible for accumulating the observable outcomes across multiple tasks to produce section- and assessment-level scores. Examples include the IRT engine used in GRE-CAT testing, the Bayesian network evidence accumulation process used in the DISC prototype, and the process of simply counting up the number of "right" answers. The measurement model in the CAF associated with a particular task specifies how this is to be accomplished.

What Happens Next: Activity Selection

This process is responsible for determining what the next task should be and when to stop the assessment. To make these decisions, adaptive assessments consult the current state of known information about a student's performance, in terms of the values of the student model variables as they have been updated by the Summary Scoring process. An instructional system will also make decisions about switching between assessment and instruction modes. Examples of activity selection processes include simple linear sequencing (the P&P GRE, although the student may choose the order in which to answer items within each section as it is administered), computerized adaptive item selection (the GRE-CAT and an operational DISC assessment), and student choice as to when to move on in a self-paced practice system (as in self-practice use of the DISC assessment capabilities).

Where Do Processes Get the Information They Need: Task/Evidence Composite Library

All four process require certain kinds of data in order to do their jobs: the Presentation Process requires the text, pictures, and other material to be displayed; Response Processing requires the "key" or other evidence rule data against which to evaluate the work products; the Summary Scoring Process requires the parameters which provide the "weights of evidence" for each task; and the Activity Selection Process requires classification and information codes used to balance the assessment form. The Task/Evidence Composite Library is a unified database that stores this information. In the P&P GRE, some of this information is used once to assemble forms, and other information is used later to score responses and accumulate evidence when completed forms are returned. In the GRE-CAT, the information must be available during testing because item selection, task scoring, and test scoring are all being carried out as testing proceeds

from one item to the next. In this respect, an operational DISC assessment would be like the GRE-CAT.

We have suggested, without detailing, the mapping between the design models built into the CAF and the four processes models. All of the design decisions made in the blueprint are reflected either directly in the implementation or in one of the processes leading up to the implementation. Again, further discussion and examples are available in Almond, Steinberg, and Mislevy (2002).

Pretesting and Calibration

In order to score an assessment, the Response Scoring Process or the Summary Scoring Process (or both) may need to build in empirical information from previous administrations of the tasks. In the case of Response Scoring, this information is incorporated into evidence rules. In the case of Summary Scoring, it appears in scoring weights or task parameters (from conditional probabilities) in the appropriate graphical models. We refer to a start-up set of data from which to estimate these values as *pretest data*, and the operation of determining the values as *calibration*.

An example for Summary Scoring occurs routinely in CAT based on IRT, such as the GRE-CAT. A collection of items may be administered to a pretest sample of students. Calibration consists of fitting the IRT model to the data, which provides estimates of item parameters that characterize the relationship of the response (the observable variable) to each item and the IRT proficiency variable, which is the single variable in the student model in such an assessment. The resulting item parameters can then be used to test future students, not only for summary scoring, but for activity selection as well. This is because the item parameters indicate how much information an item is likely to provide for a student about something that is already known from previous responses. Calibration of the DISC cases would be carried out analogously. In Bayes nets, estimation of conditional probabilities corresponds to estimation of IRT item parameters.

Conclusion

Developments in statistical modeling and estimation and new kinds of psychometric measurement models hold the promise of supporting a wider variety of educational assessments than have been traditionally used. For example, to capitalize on their potential for automated scoring, one cannot think of using them in isolation from the other processes of assessment design. All must work in concert to create an assessment that is at once coherent and practicable. Toward this end, it will be of significant benefit to have a shared framework for talking about the roles that each facet of the design elements and delivery processes play in the support of a coherent assessment argument. Evidence-centered design provides such a framework, and can thus prove useful for understanding how innovations such as cognitive modeling, new measurement models, automated scoring, and technology-based tasks fit into assessment systems.

References

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). A four-process architecture for assessment delivery, with connections to assessment design. *Journal of Technology, Learning, and Assessment 1*(5). Retrieved June 25, 2003, from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp.147-171). Urbana, IL: University of Illinois Press.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: Vol. 3. The Buros-Nebraska Symposium on measurement and testing* (, pp. 41-85). Hillsdale, NJ: Erlbaum.

Kane, M.T. (1992) An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15,* 335-374.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Appendix A

A Glossary of Evidence-centered Design Terms

Activity Selection Process. The Activity Selection Process is the part of the Assessment Cycle that selects a task or other activity for presentation to an examinee.

Administrator. The Administrator is the person responsible for setting up and maintaining the assessment. The Administrator is responsible for starting the process and configuring various choices; for example, whether or not item-level feedback will be displayed during the assessment.

Assembly Model. The Assembly Model, one of a collection of six different types of models that comprise the Conceptual Assessment Framework (CAF), provides the information required to control the selection of tasks for the creation of an assessment.

Assessment. An Assessment is a system (computer, manual, or some combination of the these) that presents examinees, or participants, with work and evaluates the results. This includes high-stakes examinations, diagnostic tests, and coached-practice systems, which include embedded assessment.

Assessment Cycle. The Assessment Cycle is composed of four basic processes: Activity Selection, Presentation, Response Processing, and Summary Scoring. The Activity Selection Process selects a task or other activity for presentation to an examinee. The Presentation Process displays the task to the examinee and captures the results (or Work Products) when the examinee performs the task. Response Processing identifies the essential features of the response and records these as a series of Observations. The Summary Scoring Process updates the scoring based on the input it receives from Response Processing. This Four-process Delivery Architecture can work in either synchronous or asynchronous mode.

Conceptual Assessment Framework (CAF). The CAF builds specific models for use in a particular assessment product (taking into account the specific purposes and requirements of that product). The CAF consists of a collection of six different types of models that define what objects are needed and how an assessment will function for a particular purpose. The models of the CAF are as follows: the Student Model, the Task Model, the Evidence Model, the Assembly Model, the Presentation Model, and the Delivery Model.

Delivery Model. The Delivery Model, one of a collection of six different types of models that comprise the CAF, describes which other models will be used, as well as other properties of the assessment that span all four processes, such as platform and security requirements.

Evaluation Rules. Evaluation Rules are a type of Evidence Rules that set the values of Observable Variables.

Evidence. In educational assessment, Evidence is information or observations that allow inferences to be made about aspects of an examinee's proficiency (which are unobservable) from evaluations of observable behaviors in given performance situations.

Evidence-centered Assessment Design (ECD). Evidence-centered Assessment Design (ECD) is a methodology for designing assessments that underscores the central role of evidentiary reasoning in assessment design. ECD is based on three premises: (1) An assessment must build around the important knowledge in the domain of interest and an understanding of how that knowledge is acquired and put to use; (2) The chain of reasoning from what participants say and do in assessments to inferences about what they know, can do, or should do next, must be based on the principles of evidentiary reasoning; (3) Purpose must be the driving force behind design decisions, which reflect constraints, resources, and conditions of use.

Evidence Model. The Evidence Model is a set of instructions for interpreting the output of a specific task. It is the bridge between the Task Model, which describes the task, and the Student Model, which describes the framework for expressing what is known about the examinee's state of knowledge. The Evidence Model generally has two parts: (1) A series of Evidence Rules that describe how to identify and characterize essential features of the Work Product, and (2) A Statistical Model that tells how the scoring should be updated given the observed features of the response.

Evidence Rules. Evidence Rules are the rubrics, algorithms, assignment functions, or other methods for evaluating the response (Work Product). They specify how values are assigned to Observable Variables and thereby identify those pieces of evidence that can be gleaned from a given response (Work Product).

Evidence Rule Data. Evidence Rule Data is data found within Response Processing. It often takes the form of logical rules.

Examinee. See *Participant*.

Examinee Record. The Examinee Record is a record of tasks to which the participant is exposed, as well as the participant's Work Products, Observables, and Scoring Record.

Four Processes. Any assessment must have four different logical processes. The four processes that comprise the Assessment Cycle include the following: (1) The Activity Selection Process—the system responsible for selecting a task from the task library; (2) The Presentation Process—the process responsible for presenting the task to the examinee; (3) Response Processing—the first step in the scoring process, which identifies the essential features of the response that provide evidence about the examinee's current knowledge, skills, and abilities; (4) The Summary Score Process—the second stage in the scoring process, which updates beliefs about the examinee's knowledge, skills, and abilities based on the evidence provided by the preceding process.

Instructions. Instructions are commands sent by the Activity Selection Process to the Presentation Process.

Measurement Model. The Measurement Model is that part of the Evidence Model that explains how the scoring should be updated given the observed features of the response.

Model. A Model is a design object in the CAF that provides requirements for one or more of the Four Processes, particularly for the data structures used by those processes (e.g., Tasks and Scoring Records). A Model describes variables, which appear in data structures used by the Four Processes, whose values are set in the course of authoring the tasks or running the assessment.

Observables/Observable Variables. Observables are variables that are produced through the application of Evidence Rules to the task Work Product. Observables describe characteristics to be evaluated in the Work Product and/or may represent aggregations of other observables.

Observation. An Observation is a specific value for an observable variable for a particular participant.

Parsing Rules. Parsing Rules are a type of Evidence Rules that reexpress the Work Product into a more "convenient" form, where convenient is interpreted to mean the form of the Work Product required by the Evaluation Rules.

Participant. A Participant is the person whose skills are being assessed. A Participant directly engages with the assessment for any of a variety of purposes (e.g., certification, tutoring, selection, drill and practice, etc.).

Platform. Platform refers to the method that will be used to deliver the presentation materials to the examinees. Platform is broadly defined to include human examiners, computer, paper and pencil, etc.

Presentation Material. Presentation Material is material that is presented to a participant as part of a task (including stimulus, rubric, prompt, possible options for multiple choice).

Presentation Material Specification. Presentation Material Specifications are a collection of specifications that describe material that will be presented to the examinee as part of a stimulus, prompt, or instructional program.

Presentation Process. The Presentation Process is the part of the Assessment Cycle that displays the task to the examinee and captures the results (or Work Products) when the examinee performs the task.

Reporting Rules. Reporting Rules describe how Student Model Variables should be combined or sampled to produce scores, and how those scores should be interpreted.

Response. See *Work Product*.

Response Processing. Response Processing is the part of the Assessment Cycle that identifies the essential features of the examinee's response and records these as a series of Observations. At one time referred to as the "Evidence Identification Process," it emphasizes the key observations in the Work Product that provide evidence.

Response Processing Data. See *Evidence Rule Data*.

Strategy. Strategy refers to the overall method that will be used to select tasks in the Assembly Model.

Student Model. The Student Model is a collection of variables representing the knowledge, skills, and abilities of an examinee about which inferences will be made. A Student Model is composed of the following types of information: (1) Student Model Variables that correspond to aspects of proficiency the assessment is meant to measure; (2) Model Type that describes the mathematical form of the Student Model (e.g., univariate IRT, multivariate IRT, or discrete Bayesian Network); (3) Reporting Rules that explain how the Student Model Variables should be combined or sampled to produce scores.

Summary Scoring Process. The Summary Scoring Process is the part of the Assessment Cycle that updates the scoring based on the input it receives from Response Processing. At one time referred to as

the "Evidence Accumulation Process," the Summary Scoring Process plays an important role in accumulating evidence.

Task. A Task is a unit of work requested from an examinee during the course of an assessment. In ECD, a task is a specific instance of a Task Model.

Task/Evidence Composite Library. The Task/Evidence Composite Library is a database of task objects along with all the information necessary to select and score them. For each such Task/Evidence Composite, the library stores (1) descriptive properties that are used to ensure content coverage and prevent overlap among tasks; (2) specific values of, or references to, Presentation Material and other environmental parameters that are used for delivering the task; (3) specific data that are used to extract the salient characteristics of Work Products; and (4) Weights of Evidence that are used to update the scoring from performances on this task, specifically, scoring weights, conditional probabilities, or parameters in a psychometric model.

Task Model. A Task Model is a generic description of a family of tasks that contains (1) a list of variables that are used to describe key features of the tasks; (2) a collection of Presentation Material Specifications that describe material that will be presented to the examinee as part of a stimulus, prompt, or instructional program; and (3) a collection of Work Product Specifications that describe the material that the task will be return to the scoring process.

Task Model Variables. Task Model Variables describe features of the task that are important for designing, calibrating, selecting, executing, and scoring it. These variables describe features of the task that are important descriptors of the task itself, such as substance, interactivity, size, and complexity, or are descriptors of the task performance environment, such as tools, help, and scaffolding.

Weights of Evidence. Weights of Evidence are parameters that provide information about the size and direction of the contribution an Observable Variable makes in updating beliefs about the state of its Student Model parent(s). The Weights of Evidence provide a way of predicting the performance of an examinee with a given state of the Student Model Variables on a given task. Examples are scoring weights in number-right scoring and item parameters in IRT models.

Work Product. A Work Product is the Examinee's response to a task from a given task model. This could be expressed as a transcript of examinee actions, an artifact created by the examinee, and/or other appropriate information. The Work Product provides an important bridge between the Task Model and the Evidence Model. In particular, Work Products are the input to the Evidence Rules.

Appendix B

Further Readings About the ECD Project

The following is an annotated list of publications that have been produced in the ECD research program. They are classified into three groups: publications about the ECD framework itself, applications of the ideas, and particular aspects of assessment design and analysis from the perspective of evidentiary reasoning.

The ECD Framework

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, *1*(5). Retrieved June 26, 2003, from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml Also available as CSE Technical Report 543. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH543.pdf [Extended discussion of the four-process delivery system architecture, including explanation of relationships between the design objects of the conceptual assessment framework and the processes and messages in an assessment delivery system.]

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2003). A framework for reusing assessment components. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.), *New developments in psychometrics* (pp. 28-288). Tokyo: Springer. [Shorter description of the four-process delivery system, with descriptions of what the four processes do and how they interact in assessments designed to achieve different purposes.]

Frase, L.T., Chudorow, M., Almond, R.G., Burstein, J., Kukich, K., Mislevy, R.J., Steinberg, L.S., & Singley, K. (2003). Technology and assessment. In H.F. O'Neil & R. Perez (Eds.), *Technology applications in assessment: A learning view* (pp. 213-244). Mahwah, NJ: Erlbaum. [This article provides an overview of developments in the use of technology in assessment. One of these is a section on the evidence-centered design system.]

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67. Forthcoming as CSE Research Report. [Currently the most comprehensive overview available of evidence-centered design, spanning assessment arguments to design elements, to delivery system architecture, and the connections within and across these levels.]

Applications

Bauer, M., Williamson, D.M., Steinberg, L.S., Mislevy, R.J., & Behrens, J.T. (April, 2001). *How to create complex measurement models: A case study of principled assessment design.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. [ECD design rationale for a simulation-based assessment of troubleshooting and design of computer networks. Foundational analysis for the NetPASS online assessment of networking skill, by the Cisco Learning Institute, Educational Testing Service, and the University of Maryland. Includes expert-novice analysis of problem-solving.]

Cameron, C.A., Beemsterboer, P.L., Johnson, L.A., Mislevy, R.J., Steinberg, L.S., & Breyer, F.J. (1999). A cognitive task analysis for dental hygiene. *Journal of Dental Education, 64,* 333-351. [Expert-novice study of expertise in problem solving in dental hygiene, with implications for assessment design.]

Mislevy, R.J., Almond, R.G., Dibello, L.V., Jenkins, F., Steinberg, L.S., Yan, D., & Senturk, D. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Technical Report 580). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TR580.pdf [Focus on estimation of conditional probability models in the Bayes net psychometric model in the Biomass prototype assessment. A fairly technical psychometric paper.]

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282. Also available as CSE Technical Report 413. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education,

UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH413.PDF [Good foundational explanation of the use of Bayesian inference in complex assessments, illustrated with the HYDRIVE intelligent tutoring system for troubleshooting aircraft hydraulics.]

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2002). Design and analysis in task-based language assessment. *Language Assessment, 19,* 477-496. Also available as CSE Technical Report 579. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TR579.pdf [ECD perspective on designing task-based language assessments. Includes examples of Bayes nets for tasks that tap multiple aspects or knowledge and skill.]

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15,* 335-374. Also available as CSE Technical Report 487. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TR487.pdf [Design and conduct of a cognitive task analysis of expertise in dental hygiene, from the perspective of informing the construction of the models in the ECD conceptual assessment framework.]

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessment. *Applied Measurement in Education, 15,* 363-378. Also available as CSE Technical Report 538. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/RML%20TR%20538.pdf [Argument that the way to design and analyze complex assessments, such as computer-based simulations, is from the perspective of the evidentiary argument—not from the perspective of technology. Ideas are illustrated in some detail with the DISC prototype assessment of problem solving in dental hygiene.]

Steinberg, L.S., & Gitomer, D.G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24,* 223-258.

[Concerns the interplay among cognitive analysis, instructional strategy, and assessment design, in the context of the HYDRIVE intelligent tutoring system for troubleshooting aircraft hydraulics.]

Steinberg, L.S., Mislevy, R.J., Almond, R.G., Baird, A.B., Cahallan, C., DiBello, L.V., Senturk, D., Yan, D., Chernick, H., & Kindfield, A.C.H. (in press). *Introduction to the Biomass project: An illustration of evidence-centered assessment design and delivery capability.* Forthcoming CSE Technical Report. Los Angeles: UCLA Center for the Study of Evaluation. [Design rationale for a standards-based, Web-delivered assessment of science inquiry, in the areas of transmission genetics and microevolution. Much discussion of working with experts and National Science Education Standards, to carry out the ECD design work and then implement a prototype assessment at the level of secondary science.]

Williamson, D. M., Bauer, M., Mislevy, R. J., & Behrens, J. T. (2003, April). *An ECD approach to designing for reusability in innovative assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., & Behrens, J. T. (2003, April). *Creating a complex measurement model using evidence-centered design*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Aspects of Assessment Design and Analysis

Almond, R.G., Herskovits, E., Mislevy, R.J., & Steinberg, L.S. (1999). Transfer of information between system and evidence models. In D. Heckerman & J. Whittaker (Eds.), *Artificial Intelligence and Statistics 99* (pp. 181-186). San Francisco: Morgan Kaufmann. Also available as CSE Technical Report 480. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH480.pdf [Concerns the technical issue of maintaining student-model and measurement-model fragments of Bayes nets, to be assembled dynamically as is required in adaptive assessments.]

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237. Also available as CSE Technical Report 434. Los Angeles: The National Center for Research on Evaluation, Standards,

Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH434.PDF [Early discussion of the kinds of variables that arise in language assessment, the roles they play in the assessment argument, and where they fit in with Bayes net modeling of performance.]

Gitomer, D.H., & Steinberg, L.S. (1999). Representational issues in assessment design. In I.E. Sigel (Ed.), *Development of mental representation* (pp. 351-370). Hillsdale, NJ: Erlbaum. [Discussion of the key role of representational forms in assessment. Addresses both the use of representational forms to provide information and elicit responses from examinees, and the role of assessments as representations themselves of what is important in a domain and how it is evaluated.]

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483. Also available as CSE Technical Report 414. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH414.PDF [Foundational, not overly technical, discussion of the role that probability-based reasoning plays in assessment and assessment design.]

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* (437-446). San Francisco: Morgan Kaufmann. Also available as CSE Technical Report 518. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH518.pdf [Discussion of Markov chain Monte Carlo estimation in a binary skills multivariate latent class model for cognitive diagnosis. Illustrated with analysis of data from Kikumi Tatsuoka's studies of mixed number subtraction.]

Mislevy, R.J., & Patz, R.J. (1995, August). On the consequences of ignoring certain conditional dependencies in cognitive diagnosis. *Proceedings of the Section on Bayesian Statistical Science: Papers presented at the Annual Meeting of the American Statistical Association,* Orlando, FL, (pp. 157-162). [Technical paper on the implications of simplifications of

Bayes net structures in assessment for computing advantage. Conclusion: Ignoring dependencies among student-model variables is generally conservative, but ignoring conditional dependencies among observations can lead to over-counting evidence.]

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97-128). Hillsdale, NJ: Erlbaum. Also available as CSE Technical Report 500. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH500.pdf

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.* New York: Teachers College Press. Also available as CSE Technical Report 534. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved June 26, 2003, from http://www.cse.ucla.edu/CRESST/Reports/newTR534.pdf [Looking from the perspective of ECD at ways that assessment can be improved by developments in statistics, technology, and cognitive psychology.]

Mislevy, R.J., Wilson, M.R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press. Forthcoming as a CSE Technical Report. [Exploration of validity, reliability, comparability, and fairness, as viewed from the perspective of evidentiary arguments.]

Williamson, D., Mislevy, R.J., & Almond, R.G. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), Uncertainty in artificial intelligence 16 (pp. 634-643). San Francisco: Morgan Kaufmann. [An initial investigation into model-fit indices for the use of Bayes nets in educational assessments.]