Evidence-Centered Assessment Design

Robert J. Mislevy, Linda S. Steinberg, & Russell G. Almond

Educational Testing Service

December 4, 1999

"Evidence-centered design" (ECD) describes a program of research and application carried out at Educational Testing Service since 1997 by Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond. The work introduces a principled framework for designing, producing, and delivering educational assessments. Special attention is accorded to assessments that incorporate features (such as complex student models and interactive simulations) that lie beyond the rules-of-thumb and analytic procedures that have evolved over the years to support familiar kinds of assessments. The contribution is not so much a particular advance in statistics, psychology, or forms of assessment as it is in laying out a coherent framework to harness recent developments of these various types toward a common purpose. The following summary gives an overview of the work.

Rationale

1. 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 et al., 1987). Off-the-shelf assessments and standardized tests are increasingly unsatisfactory for guiding learning and evaluating students' progress.

2. Advances in technology make it possible to evoke evidence of knowledge more broadly conceived, and capture more complex performances. Making sense of complex data that result is the bottleneck.

3. Advances in evidentiary reasoning (e.g., Schum, 1994) and in statistical modeling (e.g., Gelman et al., 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 that underlie familiar test theory to more varied and complex inferences from more complex data (Mislevy, 1994).

4. Advanced technologies and statistical methods aren't sufficient. 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).

5. A conceptual design framework for the elements of a coherent assessment can be laid out 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. Our design framework is based on the principles of evidentiary reasoning and the exigencies of assessment production and delivery. Designing assessment products in such a framework ensures that the way in which evidence is gathered and interpreted bears on 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 vendors, interface designers.

The resulting research program integrates and extends earlier work by the principals: Mislevy (e.g., 1994) on evidentiary reasoning in assessment, Steinberg (e.g., Steinberg & Gitomer, 1996) on cognitively-based

intelligent tutoring systems, and Almond (e.g., 1995) on building blocks for graphical probability models.

Facets of the Work

The work during 1997-1999 includes developing the assessment design framework and applying the ideas. Developing the design framework is ETS's "Portal" project. It comprises three distinguishable aspects:

- **A conceptual framework for assessment design**. The 'evidence-centered' Portal assessment design framework explicates the relationships among the inferences the assessor wants to make about the student, what needs to be observed to provide evidence for those inferences, and what features of situations evoke that evidence.

- **An object model for creating specifications for particular assessment product*s*.** The Portal object model embodies the pieces and relationships that any particular assessments needs. Instantiating instances of the objects as needed for a particular assessment ensures that an assessment will have the functionality it needs and the components will work together. This framework promotes re-usability of objects and processes.

- **Software tools for creating and managing the design objects.** The first generation of software tools to facilitate the design process have been developed. They can be used to create, manipulate, and coordinate a structured data-base containing the elements of an assessment design, which then serves as a blueprint for developing the application.

Several publications describing the conceptual model and its implications, although details of ETS's specific implementation of the approach in terms of the Portal object-model and software tools are proprietary. Various papers highlight different (interconnecting) facets of designing and implementing complex assessments, including cognitive psychology (Steinberg & Gitomer, 1996); probability-based reasoning (Almond et al., 1999); task design (Almond & Mislevy, 1999; Mislevy, Steinberg, & Almond, in press); and computer-based simulation (Mislevy et al., 1999a). Mislevy, Steinberg, Breyer, Johnson, & Almond (1999a, 1999b) apply this machinery to design a simulation-based assessment of problem-solving in dental hygiene. Almond, Steinberg, and Mislevy (1999) apply it to the challenge of defining standards for the inter-operability of assessment/instructional materials and processes.

An Overview of the Framework

Figure 1 is a schematic of the design and implementation framework. In domain analysis, designers consider the domain from a number of perspectives, including cognitive research, available curricula, expert input, standards and current testing practices, test purposes and various requirements, resources and constraints to which the proposed product might be subject. They gather information from a variety of sources and tag it in terms of key assessment design features.

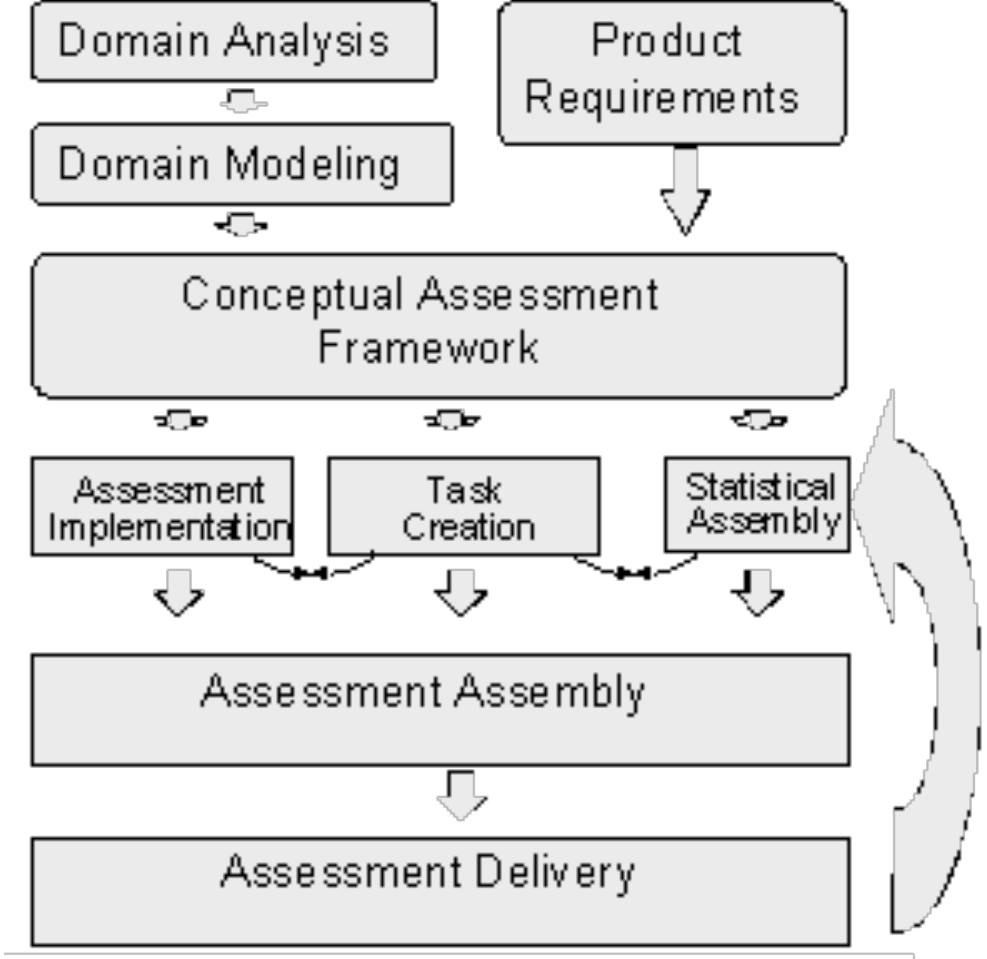

Figure 1
Schematic of Design and Implementation

In the *domain modeling* phase, the designers use information from the domain analyses to establish relationships among proficiencies, tasks, and evidence. They explore different approaches and develop high-level sketches that are consistent with what they have learned about the domain so far. They can create graphic representations and schema to convey these complex relationships, and may develop prototypes to test their assumptions.

The important design concepts are made more explicit and refined in the *conceptual assessment framework* (detailed below). The objects and specifications created here provide a blueprint for the operational aspects of work, including (a) the creation of assessments, tasks, and statistical models, (b) delivery and operation of the assessment and (c) analysis of data fed back from the field. The current version of the software tools supports the work through the creation of the CAF, and selected elements of the operational infrastructure to support (a) through (c). Specifically, we have completed the 'scoring engine' for the simulation-based assessment, and produced an architecture that supports delivery, scoring, and reporting has been for the IMS (Instructional Management Systems) inter-operability standards collaborative (Almond, Steinberg, & Mislevy, 1999).

The Conceptual Assessment Framework (CAF)

Figure 2 is a high-level schematic of the three central models in the CAF, and objects they contain. The CAF contains the core of the evidentiary-reasoning argument, from task design to observations to scoring to inferences about students.

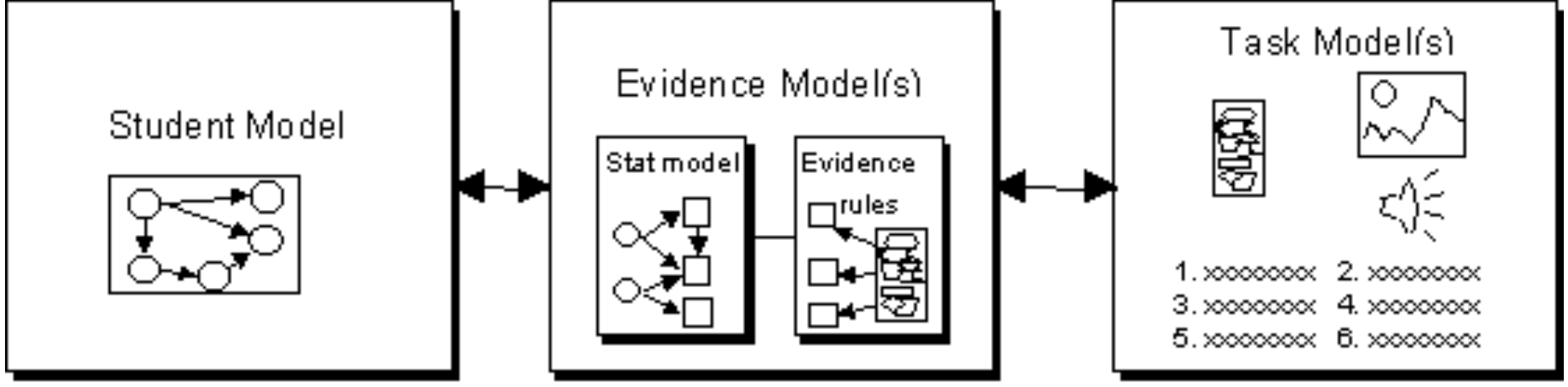

Figure 2

The Three Central Models of the CAF

The Student Model

What complex of knowledge, skills, or other attributes should be assessed? Configurations of values of student-model variables approximate selected aspects of the infinite configurations of skill and knowledge real students have, as seen from some perspective about skill and knowledge in the domain. These are the terms in which we want to determine evaluations, make decisions, or plan instruction—but we don't get to see the values directly. We see instead what students say or do, and must construe that as evidence about these student-model variables. The number and nature of the student model variables in an assessment also depend on its purpose. A single variable characterizing overall proficiency might suffice in an assessment meant only to support a pass/fail decision; a coached practice system to help students develop the same proficiency might require a finer grained student model, to monitor how a student is doing on particular aspects of skill and knowledge for which we can offer feedback.

The student model in Figure 2 depicts student model variables as circles, while the arrows represent important empirical or theoretical associations. We use a statistical model to manage our knowledge about a given student's unobservable values for these variables at any given point in time, expressing it as a probability distribution that can be updated in light of new evidence. In particular, the student model takes the form of a fragment of a Bayesian inference network, or Bayes net (Spiegelhalter et al., 1993).

Evidence Models

What behaviors or performances should reveal those constructs, and what is the connection? An evidence model lays out an argument about why and how our observations in a given task situation constitute evidence about student model variables.

Figure 2 shows there are two parts to the evidence model. The *evaluative submodel* concerns extracting the salient features of whatever the student says, does, or creates in the task situation—the "work product" represented by the jumble of shapes at the far right of the evidence model. It is a unique human production, as simple as a response to a multiple-choice item or as complex as evaluation and treatment cycles in a patient-management problem. Three squares coming out of the work product. represent "observable variables," evaluative summaries of whatever the designer has determined are the key aspects of the performance in light of the assessment's purpose. Evaluation rules map unique human actions into a common interpretative framework, effectively laying out the argument about what is important in a performance. These rules can be as simple as determining whether the response to a multiple-choice item is correct or as complex as an expert's holistic evaluation of four aspects of an unconstrained patient-management solution. They can be automated, demand human judgment, or require both in combination.

The *statistical submodel* of the evidence model expresses the how the observable variables depend, in probability, on student model variables. This is effectively the argument for synthesizing evidence across multiple tasks or from different performances. Figure 2 shows that the observables are modeled as depending on some subset of the student model variables. Familiar models from test theory, such as item response theory and latent class models, are examples of statistical models in which values of observed variables depend probabilistically on values of unobservable variables. We can express these familiar models as special cases of Bayes nets, and extend the ideas as appropriate to the nature of the student model and observable variables.

Task Models

What tasks or situations should elicit those behaviors? A task model provides a framework for constructing and describing the situations in which examinees act. Task model variables play many roles, including structuring task construction, focusing the evidentiary value of tasks, guiding assessment assembly, implicitly defining student-model variables, and conditioning the statistical argument between observations and student-model variables (Mislevy, Steinberg, & Almond, in press). A task model includes specifications for the environment in which the student will say, do, or produce something; for example, characteristics of stimulus material, instructions, help, tools, affordances. It also includes specifications for the work product, the form in which what the student says, does, or produces will be captured.

Implementation and Delivery

Figure 3 sketches four principle processes that take place in an assessment. Some are compressed or implicit in familiar forms of assessment. Explicating them makes it easier to design re-usable, inter-operable, components. The Activity Selection Process selects a task (or other activity) and instructs the Presentation Process to display it. When the examinee has finished interacting with the item, then the Presentation Process sends the results (a Work Product) to the Evidence Identification Process. This process identifies key Observations about the results and passes them to the Evidence Accumulation Process which updates the Examinee record. The Activity Selection then makes a decision about what to do next based on the current beliefs about the examinee.

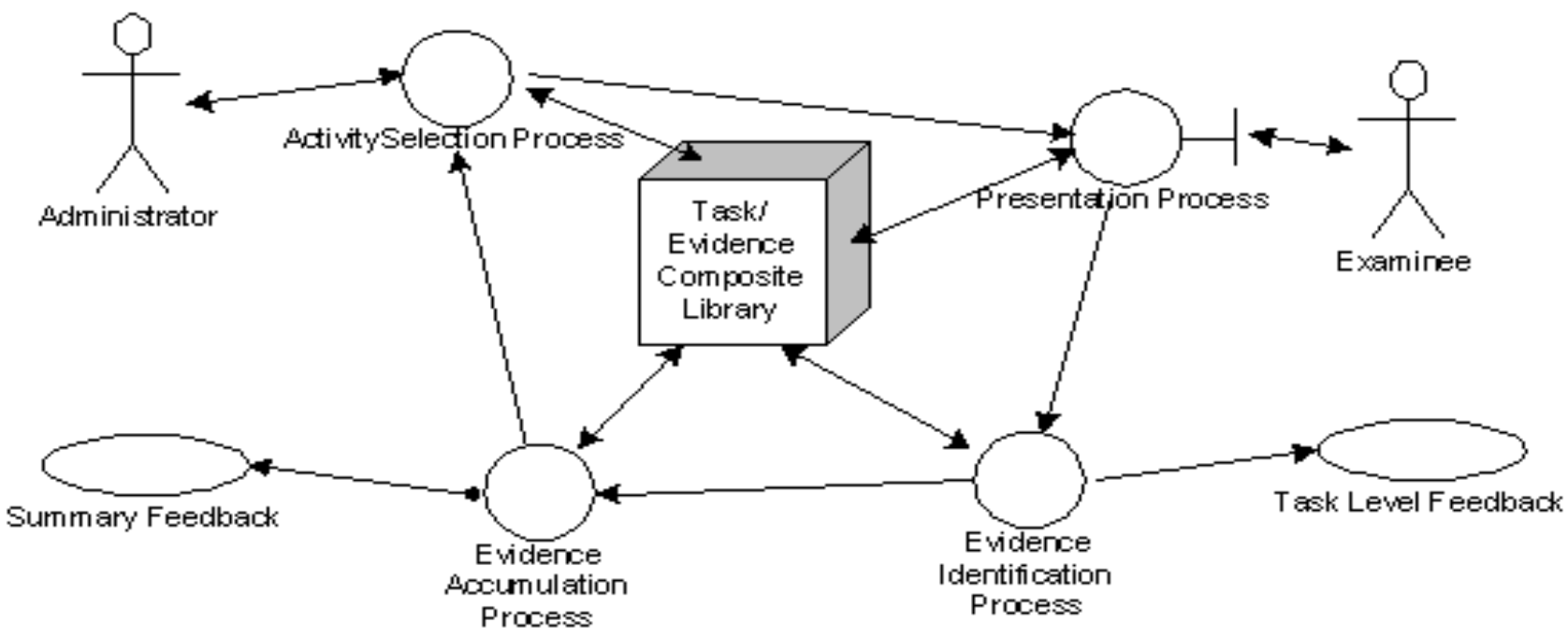

Figure 3

Processes in the Assessement Cycle

Ensuring that these processes interact coherently requires standards for the messages they must pass from one to another (Figure 4). The framework for defining the forms and the contents of the messages in a given assessment—importantly, not the forms or the content themselves—are specified in the evidence-centered object model, in the CAF in particular. In this way, designing an assessment within the common evidence-centered framework ensures the coordination of operational processes. Analogously, fully specifying CAF objects helps the assessment designer lay out specifications for task creation and statistical analyses.

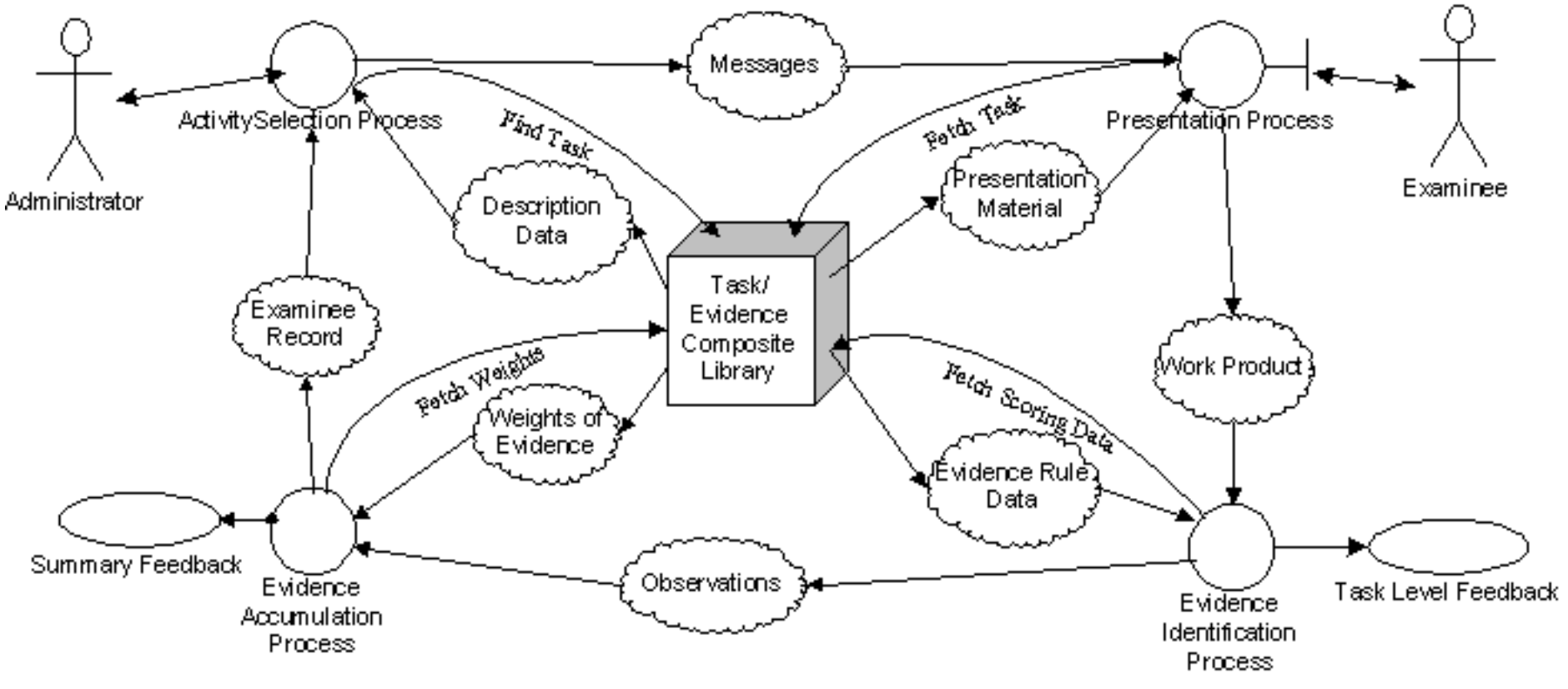

Figure 4

A More Detailed View of Processes, Showing Messages between Processes

Importance of the Work

Opportunities borne of new technologies, desires borne of new understandings of learning—a new generation of assessment beckons. To realize the vision, we must reconceive how we think about assessment, from purposes and designs to production and delivery. Our approach will need to link assessment designs more explicitly to the inferences we want to draw on the one hand, and to the processes we need to create and deliver them. Providing these links is raison d'être of the work on evidence-centered assessment design outlined here. With a framework such as the one described here, can we hope to provide a flexible framework for developing wide ranges of assessments for many purposes, while ensuring the coherence of contributions of different bodies of expertise.

References

Almond, R.G. (1995). *Graphical belief modelling*. London: Chapman and Hall.

Almond, R.G., Herskovits, E., Mislevy, R.J., and 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.

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237.

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (1999). *A sample assessment using the four process framework*. White paper prepared for the IMS Inter-Operability Standards Working Group. Princeton, NJ: Educational Testing Service.

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: The Buros-Nebraska Symposium on measurement and testing* (Vol. 3) (pp. 41-85). Hillsdale, NJ: Erlbaum.

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., 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 Publishers, Inc.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice*. Hillsdale, NJ: Erlbaum.

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

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999b). Making sense of data from complex assessments. Paper presented at the 1999 CRESST Conference, Los Angeles, CA, September 16-17, 1999. 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.

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

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., & Cowell, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8,* 219-283.