

Using Evidence Models to Aggregate Information

Russell G. Almond,¹

Author Contact: ralmond@ets.org

Date: 2002/12/13 18:47:06

ABSTRACT

The major challenge in establishing a marketplace of interoperable learning objects is determining how to update the Proficiency Model of the Learning Management System on the basis of the outcomes from a Learning Object. An *Evidence Model* which answers the question “How does the outcome from the Learning Object provide evidence for the proficiency I wish to measure?” offers a sound theoretical foundation for aggregating information from multiple sources. Automating the construction of Evidence Models poses some interesting ontological and statistical challenges, however, it would solve a fundamental problems in both Education and any other field which needs to integrate information across many sources.

Information Aggregation in Electronic Learning

Already many vendors make diverse, innovative kinds of on-line training and educational material. The challenge lies in getting them to talk to each other. Many organizations have looked at the problem of standardizing the communication between learning objects, e.g., IMS, SIF, OKI, ADL. Perhaps nowhere is the vision more clearly expressed than in ADL’s *Scorable Content Object Reference Model (SCORM)*. SCORM envisions a world in which a Learning Management System (LMS) can select a Content Object, start the learner interacting with that object, and then be able to aggregate the results produced with results generated by other Content Objects.

The question is how to do that aggregation. In general, the LMS and Content Object will be designed by different groups with different educational objectives. What does a given result from a particular Content Object mean in terms of the goals of the course being managed by the LMS? The problem is further complicated as both the outcomes from the Content Object and the internal proficiency model of the LMS could be multivariate (as is necessary to support both complex simulations and diagnostic reporting).

ETS has developed a methodology called “Evidence Centered Design” (Mislevy, Steinberg, and Almond [2002b]) which includes two tools which can provide perspective on this problem. The first is a generic Four Process Architecture for assessments (Almond, Steinberg, and Mislevy [2002]). The second is a dynamic model for aggregating evidence from disparate sources (Almond and Mislevy [1999]). The key is the *Evidence Model*—a model built for each Content Object which tells how to combine the results for that object with the general model for proficiency found in the LMS.

The Four Process Architecture

Almond, Steinberg and Mislevy [2002b] lay out an ideal architecture of an assessment which consists of four processes (Figure 1):

- The *Activity Selection Process* determines what task is assigned next and when to stop assigning tasks. It can use intermediate outcomes from the Summary Scoring Process when making these decisions.
- The *Presentation Process* presents the task to the learner and gathers the resulting response. The response could be as simple as the index of a selected response or a complex record of a performance.
- The *Response Processing* parses and interprets the raw response producing a task level outcome. For example, it decides whether a selection was correct or incorrect or assigns a grade to an essay.
- The *Summary Scoring Process* aggregates the evidence from individual tasks. It updates an internal Proficiency Model for each learner on the basis of the evidence obtained so far.

Two features of this architecture are important when looking at future applications. First, the communications channels must handle complex, multivariate data. Second, Response Processing is logically

¹ Educational Testing Service

Using Evidence Models to Aggregate Information

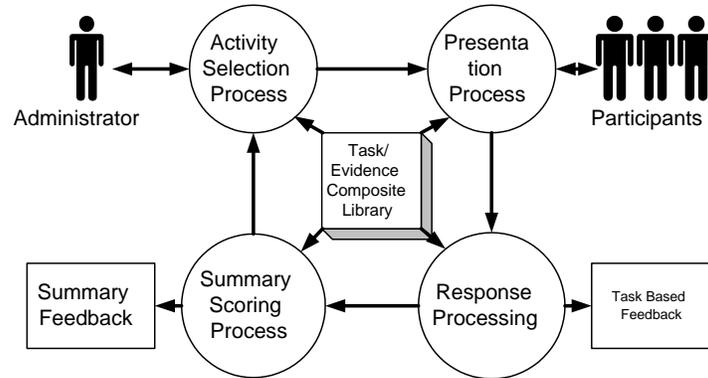


Figure 1. The Four Process Architecture for Educational Assessments

separate from both Presentation and Summary Scoring. Swapping Response Processing modules allows a system to adapt the outcomes to the correct level of detail for the internal Proficiency Model of the Summary Scoring Process.

Although originally designed for the microcosm of a single assessment, the Four Process Architecture works equally well for describing the relationship between a Learning Management System and a collection of Content Objects. In this situation, each Content Object is a Presentation Process. Response Processing plays a particularly vital role in this context, adapting the number and type of the results from the Content Objects to match what the LMS requires. For example, it could apply a cut score to produce a binary response, it could remap a letter grade to a numeric value, or it could summarize across multiple results.

The Evidence Model as an Adapter

Evidence Centered Design (Mislevy, Steinberg and Almond [2002b]) provides a conceptual framework for describing an assessment based on four kinds of models (Figure 2).

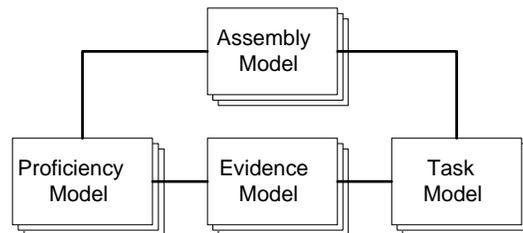


Figure 2. Four Models of the Evidence Centered Design Conceptual Assessment Framework

- The *Proficiency Model* describes what kinds of claims we wish to make about a learner and what knowledges, skills, and abilities are necessary to support those claims.
- The *Assembly Model* describes the mixtures of tasks in an assessment, and rules for ensuring that enough information is gathered to make reliable inferences about the claims.
- The *Task Models* describe characteristics of families of tasks, including what kind of material they use, what kind of responses they produce, and what kind of descriptive classification information (meta-data) is available about the tasks from that family.
- The *Evidence Models* describe how to update the learner's Proficiency Model based on the responses from tasks from a given Task Model. It describes both how to process the raw responses to obtain observed outcomes (Response Processing) and how to update proficiency variables based on the observed outcomes (Summary Scoring).

Using Evidence Models to Aggregate Information

To be able to use an arbitrary Content Object with a given Learning Management system, we must produce specifications for the Response Processing and Summary Scoring associated with that task. In general, the LMS will provide the Proficiency and Assembly Models, and the Content Object will provide the Task Model. The Evidence Model, which provides the bridge between the Task and Proficiency Model, must be crafted for each new type of Content Object. The challenge is to find a way to build those Evidence Models efficiently.

The Bayesian Statistical Paradigm offers a sound approach to designing Proficiency and Evidence Models (Almond and Mislevy [1999]). The heart of the Proficiency Model is a probability distribution over the variables representing our state of knowledge about the learner’s proficiencies.

The half of the Evidence Model model used in Summary Scoring is a predictive distribution for the observed outcomes given the learner’s proficiency variables. (The half used in Response Processing is the rules for computing the values of the outcome variables.) The Summary Scoring process uses Bayes theorem to update our beliefs about the proficiency variables based on the observed outcomes.

Because, in general, there can be multiple proficiency variables and observed outcomes, we use graphical models to represent both Proficiency and Evidence Models. Figure 3 shows a typical Proficiency Model and Evidence Models for four different kinds of tasks. The Summary Scoring process chooses an Evidence Model appropriate for the task, “docks” it with the Proficiency Model, updates based on the observations, and then discards the now used Evidence Model. It can also run this process backwards to make predictions about the amount of information which could be obtained from a given task.

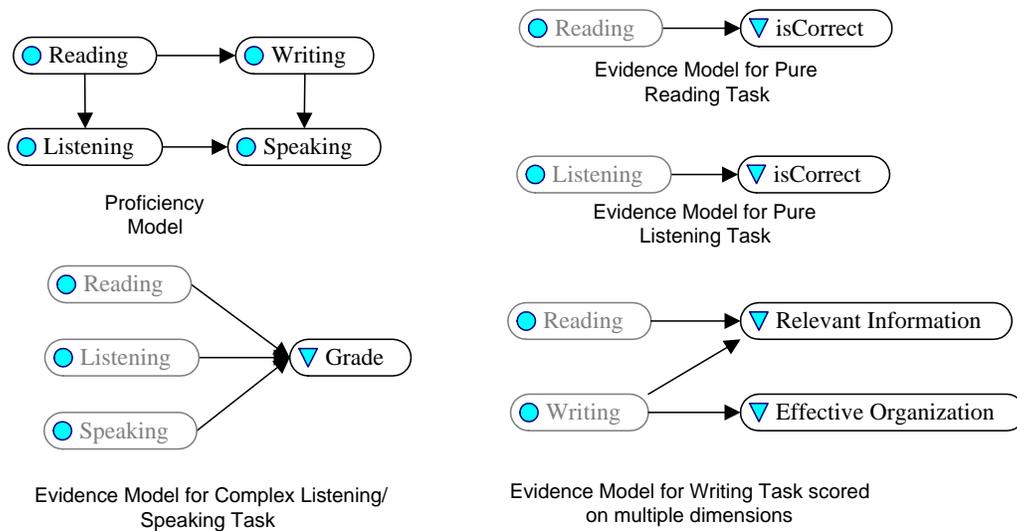


Figure 3. Proficiency Model and Evidence Models from Several Tasks

A Proficiency Model and several Evidence Models which might be used in a language assessment. Note that variables shown in gray in the Evidence Models are references to variables in the Proficiency Model.

Challenge I: Intelligent Matching

The Almond and Mislevy [1999] model was inspired by attempting to apply the ideas of knowledge-based model construction (Breese, Goldman and Wellman [1994]) to educational testing. The vision is to assemble a valid statistical model for a problem using conventional Artificial Intelligence techniques like production rules and case based reasoning. This provides a way of making a model focused on the problem at hand which can be supported without unreasonable computational expense.

This could be implemented with today’s technology, but only by handcrafting an Evidence Model for each pair of Content Object and LMS to be used together (but Gertner, Conata and VanLehn, 1998, achieve some automation in a restricted domain). The process has two labor intensive steps: linking the observable

Using Evidence Models to Aggregate Information

outcome variables to the proficiency variables, and determining the strength of the relationships of those variables. Currently both of these tasks require a collaborative effort between psychometricians and subject matter experts.

The linking step itself has two parts. The first is making connections between the proficiency variables and the observed outcomes. This should be based classification meta-data for both the LMS and the Content Object. However, the same term can mean very different things when used in different places. For example, two Content Objects on the “Civil War” would refer to different time periods if one was produced in England and one in the United States. To address this problem, research and standardization of various ontological schemes, in particular, applied research putting these ontologies into service of the model construction problem, will be necessary. Building these ontologies will be a large, expensive effort requiring the collaboration of many organizations.

The second part of the linking step is defining the observable outcome variables for each Content Object. Recall that these can be arbitrary transformations or combinations of the results from the Content Object. Again the Content Object meta-data should form the basis for these transformations, in particular, information about the scale and reliability of each of the results it produces (Hand [1993]).

Challenge II: Assessing Relationship Strength

The second step in building the Evidence Model is assigning numeric values to the strengths of the relationships. In the Bayesian paradigm these are the probabilities of seeing the potential outcomes given a particular profile of proficiency variables. Ideally, these would come from empirical observation, but that would require large linking studies involving learners with a variety of known profiles. Thus, they may be prohibitively expensive in practice.

A second source for the numeric values is expert opinion. However, it is difficult to get experts properly calibrated to make probabilistic judgements (Kahneman, Slovic and Tversky [1982]). We need to take better advantage of collateral information about the tasks in the Content Object to assist with this task. Mislevy, Steinberg and Almond [2002a] lay out some ways in which Content Object meta-data (task feature variables) can be used to help determine Evidence Model parameters for a family of related Content Objects. The first approach (following Embretson [1993]) uses the task feature variables in a generalized regression model. The second approach uses natural hierarchies of related tasks to produce hierarchical Bayesian models where Evidence Models are assigned parameter values on the basis of similar Content Objects (Almond, Mislevy, and Steinberg [1997], Glas and van der Linden [2001]).

In any case, such methods are still in early development. In many cases, the potential collateral information will need to be specified using controlled and standardized vocabularies. Thus, this step presumes that progress has been made on the ontological issues described in the previous section.

The Vision

The vision described here is achievable. We can do it now, albeit laboriously and by hand. It will take many attempts at doing this by hand and documenting the process before we understand it well enough to automate it. Furthermore, many parts of the automation will require large shared ontologies which will take many individuals and organizations to build.

As a result of this research effort, we will have achieved true “plug and play” technology. Learning objects will be self-describing entities which will contain the knowledge necessary to forge the links between many different kinds of learning object.

However, the paradigm described here is not limited to education. It could be used in a number of domains in which integrating information from diverse sources is necessary. It has been applied in modelling management decision making (Bradshaw and Boose [1990]), military intelligence (Laskey and Mahoney[1997], Koller and Pfeffer[1997]), and reliability engineering (Almond, Bradshaw and Madigan [1994]). Thus it has real potential to change the overwhelming amount of data we receive from a multitude of sources and turn it into evidence for or against claims we care about.

Using Evidence Models to Aggregate Information

Referenced Papers

- Almond, R.G., J.M. Bradshaw, and D. Madigan [1994].** "Reuse and Sharing of Graphical Belief Network Components." in P. Cheeseman and W. Oldford (eds.) *Selecting Models from Data: Artificial Intelligence and Statistics IV*, Springer-Verlag, 113–122.
- Almond, R.G., R.J. Mislevy, and L. Steinberg [1997].** "Using Prototype-Instance Hierarchies to model Global Dependence." *AMS Summer Research Conference on Graphical Markov Models, Influence Diagrams, and Bayesian Belief Networks* Seattle, Washington.
- Almond, R.G., and R.J. Mislevy. [1999].** "Graphical models and computerized adaptive testing." *Applied Psychological Measurement*. **23** 223–238.
- Almond, R.G., L.S. Steinberg, and R.J. Mislevy [2002].** "Enhancing the design and delivery of assessment systems: A four-process architecture." *Journal of Technology, Learning, and Assessment*, **1** (5). Available from <http://www.jtla.org>.
- Bradshaw, J.M., and J.H. Boose [1990].** "Decision Analysis Techniques for Knowledge Acquisition: Combining Information and Preferences using Acquinas and Axot1." *International Journal of Man-Machine Studies*, **32**: 121–186.
- Breese, J.S., R.P. Goldman, and M.P. Wellman [1994].** "Introduction to the Special Section on Knowledge-Based Construction of Probabilistic and Decision Models." *IEEE Transactions on Systems, Man, and Cybernetics*, **24**, 1577–1579.
- Embretson, S.E. [1993].** "Psychometric models for learning and cognitive processes." In N. Frederiksen, R.J. Mislevy and I.I. Bejar (Eds.) *Test Theory for a New Generation of Tests*. Laurence Erlbaum Associates. 125–150.
- Gertner, A., Conati, C., and VanLehn, K. [1998].** "Procedural help in Andes: Generating hints using a Bayesian network student model." In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence AAAI-98*. The MIT Press. 106-111.
- Glas, C.A.W. and W.J. van der Linden [2001].** "Modeling Variability in Item Parameters in CAT." *International Meeting of the Psychometrics Society* July, 2001. Osaka, Japan.
- Hand, D.J. [1993].** "Measurement scales as metadata", in Hand, D.J. (ed.) *Artificial Intelligence Frontiers in Statistics: AI and Statistics III*. Chapman and Hall, 54–64.
- Kahneman, D., P. Slovic, and A. Tversky [1982].** *Judgement under uncertainty: Heuristics and biases*. Cambridge University Press.
- Koller, D. and A. Pfeffer [1997].** "Object-Oriented Bayesian Networks." *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97)*. Morgan Kaufmann. 302–313.
- Laskey, K.B. and S.M. Mahoney [1997].** "Network fragments: Representing knowledge for constructing probabilistic models." *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97)*. Morgan Kaufmann. 334–341.
- Mislevy, R.J., L.S. Steinberg, and R.G. Almond [2002a].** "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, 97–128.
- Mislevy, R.J., L.S. Steinberg, and R.G. Almond [2002b].** "On the structure of educational assessment (with Discussion)." *Measurement: Interdisciplinary Research and Perspective*. **1**, 3–67.

Referenced Organizations

- ADL.** Advanced Distributed Learning Initiative. <http://www.adlnet.org/>
- IMS.** IMS Global Learning Consortium, Inc. <http://www.imsglobal.org/>
- OKI.** Open Knowledge Initiative, Massachusetts Institute of Technology. <http://web.mit.edu/oki/>
- SIF.** Schools Interoperability Framework. <http://www.sifinfo.org/>