**Testing Tests**

*Russell Almond, David Williamson, Duanli Yan*
*Educational Testing Service*

**Abstract**

Almond and Mislevy [1999] lays out a general framework for educational assessment based on graphical models, particularly Bayesian networks. Actually designing and building such an assessment for a particular purpose, for example, a state assessment for annual progress under the No Child Left Behind Act, is a complex process consisting of many phases: Analysis, Design, Implementation, Pretesting, and Deployment (Mislevy, Steinberg, and Almond [2003] lay out the stages of this process). The resulting model (partitioned into many pieces) is a combination of expert opinion and pretest data. However, it is also constrained by a set of requirements based on the purpose to which the scores generated by this assessment will be put.

An important part of the acceptance of the assessment by the administrating authority is that it be thoroughly tested. As in any engineering process, there are opportunities for testing throughout the process (Stark[1992]). We propose to explore some of the methods for testing throughout the process:

1. **Analysis**—Gathering initial requirements and making preliminary drafts of the key design models for scoring and task authoring.
   - Building prototype score reports to check all key requirements are satisfied.
2. **Design**—Fleshing out the preliminary drafts to make specifications for tasks and scoring models.
   - Running prototypical students through the scoring model to check that the scores are reasonable.
   - Scoring simulated responses to make sure the model correctly classifies students who behave according to the idea model.
   - Simulating item selection algorithms to make sure the model and selection algorithm interact reasonably.
3. **Implementation**—Authoring tasks and writing software for delivery and scoring.
   - Simulating item selection algorithms to make sure constraints on test forms are met.
   - Unit testing of scoring engine.
   - Unit testing of calibration procedure. This is difficult because MCMC procedure uses random numbers.
4. **Pilot testing**—Pretesting with preliminary data given to students like typical examinees.
   - Calibrate items according to pretest data.
   - Check model fit; look at item and model improvement.
   - Check assessment for fairness; look for conditional dependence on undesirable variables (e.g., race and gender of student).
5. **Deployment**—Putting an operational test program out in the field.
   - Validity studies showing that scores are properly correlated with requirements.

**References**

**Almond, R.G., & Mislevy, R.J. [1999].**. "Graphical models and computerized adaptive testing." *Applied Psychological Measurement.* **23** 223–238.

**Mislevy, R.J., Steinberg, L.S., & Almond, R.G.[2003]**. "On the Structure of Educational Assessments." *Measurement: Interdisciplinary Research and Perspectives.* **1**, 3–62.

**Stark, J. [1992]**. *Engineering Information Management Systems: Beyond CAD/CAM to Concurrent Engineering Support.* Van Nostrand Reinhold.