

**A PROPOSAL FOR A CRITERION-  
REFERENCED EVALUATION SYSTEM  
FOR ENGLISH SKILLS ACQUISITION**

G. DeMauro

Overview

The implementation of the New York State Learning Standards has helped operationalize the level of English skills needed for adequate performance in the monolingual English instructional environment. Until now, the goal was to broadly define a level of English language skills that distinguished whether the optimal environment was transitional bilingual or monolingual English. A major variable confounding that decision was the nature of the instruction. That is, while there was a need for the State to determine where the child would best succeed, this determination was made with only a general description of the curricula of the two types of classrooms: Perhaps the child would do better in the monolingual mathematics classroom but the bilingual social studies classroom, or perhaps the child would do better in the monolingual mathematics classroom in one small city but better in the bilingual mathematics classroom in a small rural town.

The Learning Standards, by taking a giant step forward in making the educational goals explicit, has given educators the opportunity to think about the level of English needed to achieve certain criteria. While this is not uniform from locale to locale or grade to grade, it does compel the school districts to broad uniformity. The students are expected, by fourth grade, to produce correct English writing that is relatively independent of the context. Mathematics

Demands problem solving techniques that must be planned and not only demonstrated by the teacher, but communicated back to the teacher.

It remains, then, to find not just a score, but to define the domain of English skills demanded in the monolingual classroom. This can be accomplished in different ways. The following is a technical procedure for beginning this definition.

### Test Development

The test under development is best conceived as an item pool that spans the range of little or no English to English proficiency. It also spans the grade ranges. For the appropriate level of skills and the appropriate developmental level, this pool could be the source of test forms that overlap across proficiency levels and developmental or grade levels. We propose to use the State model for test development to the extent possible.

Particular attention should be given to the skills needed to perform in the monolingual English classroom in which the Learning Standards are successfully implemented.

### Field Testing

Because the model seeks to chart development toward the acquisition of the skills needed in the monolingual classroom, the item pool should be field-tested on both monolingual curriculum students and bilingual curriculum students as well as students receiving English as a Second Language services without a formal bilingual program.

Items should be calibrated onto a single scale using IRT models. Basically, this means that the difficulty of the test questions and the scores of the students should all be gauged to the same scale. If a student has a higher score value than a given item, this means that the student has a

greater than 50% chance of answering that item correctly. If a student has a lower score value than a given item, then the probability drops below 50%.

Each item should be carefully classified according to the skill it is intended to measure. Factor analytic studies should be used to describe and confirm the construct being measured. Finally, as part of the calibration, all items should undergo Differential Item Functioning (DIF) analyses to assure that they are measuring the same construct for the monolingual curriculum students and for the English Language Learners. There should be about 1200 students tested per item.

### Classification

Techniques should be applied to find the scores that maximize the distinction not only between the two populations, but also between levels of skills within each. In particular, there should be a classification of monolingual curriculum students in terms of their achievement of the Learning Standards, either by their school district, or student by student. For example, an environment that is defined as successfully implementing the Learning Standards according to test scores, etc., would be given a different classification than a similar environment that does not have the same level of success. This requires that the committees begin to define criteria for sampling that includes variables related to implementation of the Learning Standards.

Classificatory discriminant analysis, then, can be used to determine scores that would best distinguish these various populations.

Because the items are associated with skills, and the items and the student scores have been placed on the same scale, it is possible to compare every student's score to the criterion score. This would identify the English language deficits of each student in terms of the demands of the

monolingual English classroom where the Learning Standards are achieved. Very simply, between every score and the criterion score would be test items that are identified for what they measure. This could then be translated into skills needed to achieve the criterion score or skills needed to achieve the Learning Standards in a monolingual classroom.

### Equating

The item pool, because it is all calibrated on the same scale, would be vertically equated. That is, field test forms could be developed with sufficient overlap to enable single scaling from grade to grade. Some items would be put aside to be used to anchor new items in field tests every year to continually replenish the calibrated item pool.

The single scale design enables not only comparison of performance from year to year, but also comparison in terms of the requisite skills needed each year. District, school, and program means could be charted over time to determine strengths and weaknesses.