TO: Alberto Sorongon, Karen Tourangeau, Christine Nord, WESTAT
FROM: Michelle Najarian, ETS
DATE: February 3, 2010
SUBJECT: Recommendation for Administration of the Spanish Basic Reading Skills Assessment in the ECLS-K:11

For the ECLS-K:11, there is a desire to have a measure of Spanish-speaking children's reading skills and knowledge in their native language. For those children who are identified as Spanish-speakers from school records, and who do not pass the language screener, Spanish reading ability will be assessed by administering a translation of some items from the basic skills and vocabulary categories of the reading assessment in English, that are anticipated to be fairly easy to translate into Spanish and maintain similar psychometric properties.

This would result in a Spanish Basic Reading Skills assessment, comprising letter identification, letter sounds, print familiarity, and simple vocabulary items that may equate to sub-scores of the English version of the reading test. These items, if psychometrically similar in the English and Spanish versions, would then facilitate the analyses of basic reading skills for all English- and Spanish- speaking children in their "preferred language," as well as support analyses comparing Spanish speakers' basic skills in English and Spanish. The administration of the SBRS will be entirely in Spanish, including the assessor instructions and child responses.

One purpose of the Spanish field test in fall 2009 was to identify items in English and Spanish that equate, thus permitting a combined calibration in basic reading skills in both English and Spanish. The purpose of this memo is to provide a recommendation for administering the Spanish Basic Reading Skills (SBRS) assessment in the ECLS-K:11. The types of analyses will be discussed, followed by the analysis results, recommendation, and next steps.

Analysis

We used two approaches in the initial analysis of the field test item data for the SBRS: classical item analysis and Item Response Theory (IRT), both providing information on item difficulty and the relationship of individual items to the construct as a whole

Classical item analysis includes the percent correct (P+) for each item, the correlation of performance on each item to performance on the test as a whole (r-biserial), omit rates, and distracter analysis (number and overall test performance for children choosing each response option for multiple choice items, or for children answering right/wrong for open ended items), as well as the internal consistency (alpha coefficient) for the set of items. Strengths of item analysis include the possibility of identifying weak distracters (i.e., options chosen by very few test takers), multiple keys (more than one correct or nearly correct answer), or items omitted by unusually high numbers of children. It also makes it possible to observe whether a response option designated as incorrect may be chosen by test takers who scored as high or higher, on average, on the test as a whole than did the group choosing the intended correct answer. This situation would result in a low or negative r-biserial, that could be an indicator of confusing or ambiguous language or presentation that may be causing children to misunderstand the intent of the question, or a different and equally correct interpretation not anticipated by the item writer.

A limitation of classical item analysis is that statistics are distorted by treatment of omits, whether they are included or excluded from the denominator of the percent correct. *Including* omits implicitly makes the assumption that ALL children who omitted an item would have gotten it wrong if they had attempted it, while *excluding* omits from P+ is equivalent to assuming that the same proportion of children who omitted would have given a correct answer. Neither assumption is likely to be true. Even if there are negligible numbers of omits, percent correct does not tell the whole story of item difficulty: all other things being equal, a multiple choice item will tend to have a higher percent correct than a comparable open ended item because of the possibility of guessing. R-biserials provide a convenient measure of the strength of the relationship of each item to the total

construct, but they too are affected by omits. The P+ and r-biserial reported in classical item analysis contribute to an overall evaluation of a test item, but may be attenuated for items that perform well at one level of ability but not another.

Even with its limitations, however, classical item analysis provides an overview of the item quality and overall assessment consistency as well as details. It provides a look at the functioning of each individual response option, as opposed to IRT, which only considers whether the correct option was chosen. And when coupled with IRT analysis, which accounts for omits and the possibility of guessing, and shows the ability levels at which an item performs well, the result is a comprehensive assortment of statistics for item and assessment evaluation.

PARSCALE is the IRT program used for calibrating test takers' ability levels on a common scale regardless of the assortment of items administered. The graphs (item characteristic curves) generated in conjunction with PARSCALE are a visual representation of the fit of the IRT model to the data. The IRT difficulty parameter ("b") for each item is on the same scale as the ability estimate (theta) for each child, allowing for matching a set of test items to the range of ability of sampled children. The IRT ("a") parameter, "discrimination" is analogous to the r-biserial of classical item analysis. IRT output includes a graph visually illustrating performance on the item for children at different points across the range of ability.

We evaluated item quality and potential for use in the SBRS by reviewing all of the available information for each item, including:

- Difficulty: Matching the difficulty of the test questions to represent items measuring basic reading skills.
- Common Functionality across Languages: Evaluating the functioning of the items in both English and Spanish to determine if they are equatable and thus can be used in a combined calibration.
- Test Specifications: Targeting percentages of the basic skills and vocabulary content categories from the framework.
- Psychometric Characteristics: Selecting items that do a good job of discriminating among achievement levels.
- Linking: Having sufficient overlap of items across subsequent grade levels so that a stable scale can be established for measuring status and gain, as well as having an adequate number of items carried over from the ECLS-K to permit cross-cohort comparisons.
- Assessor Feedback: Considering observations made by the field staff on the item functioning.
- Measurement of Gain: Evaluating whether performance improved in subsequent years.

For example, an item with P+ = .90 (90% of children answering correctly) and IRT b=-2.0 would appear to be an easy item, potentially suitable for a test of basic reading skills. But a low r-biserial (below about 0.30) or a relatively flat IRT "a" parameter (below 0.50 or so), suggest a weak relationship between the item and the test as a whole. In other words, while the item is somewhat easy, it is not useful in differentiating different levels of basic reading skills because it is about equally difficult for low-ability and high-ability students.

Pooling Data and Samples

In order to measure each child's status accurately, it is important that each child receive a set of test items that is appropriate to his or her skill level. Selection of potential items brings together two sets of information: the difficulty parameters for each of the items in the pool, and the range of ability expected in the each round. Calibration of these two pieces of information *on the same scale*, so that they may be used in conjunction with each other, was accomplished by means of Item Response Theory (IRT) analysis. (In-depth discussions of the application of IRT to longitudinal studies may be found in the ECLS-K psychometric reports.)

IRT calibration was carried out by *pooling the following datasets together*:[1]

- ECLS-K:11 fall kindergarten Spanish field test (approximately 1100 cases)
- ECLS-K:11 fall kindergarten field test (approximately 450 cases)
- ECLS-K:11 fall first grade field test (approximately 400 cases)
- ECLS-K fall kindergarten national - items selected from general knowledge assessment (approximately 18,000 cases)
- ECLS-K spring kindergarten national - items selected from general knowledge assessment (approximately 19,000 cases)
- ECLS-K fall first grade national - items selected from general knowledge assessment (data collected only for a subsample of about 5,000)
- ECLS-K spring first grade national - items selected from general knowledge assessment (approximately 16,000 cases)

The overlapping items shared by two or more datasets serve as an anchor, so that parameters for items and samples from different assessments are all on a common scale. Data from the ECLS-K:11 English and Spanish field test samples supply a link between the newly-developed field test items and the ECLS-K kindergarten/first grade assessments, enabling them to be put on the common scale necessary for direct comparisons. The large samples from the ECLS-K data collections also serve to stabilize parameter estimates that would be unreliable if only the small samples of the fall 2009 field test was available.

Pooling the datasets together also provides estimated values for the mean ability levels for each dataset on the same scale. Although the datasets are pooled together, the samples are treated separately to preserve the ability ranges of each dataset. Mean ability levels for each of the datasets above were calculated from the pooled sample.

Results

Review of the data from the Spanish field test showed that many of the SBRS items worked well as a measure of basic reading skills, *and* were common-functioning with the English version of the items. Figure 1 illustrates the behavior observed for most items.[2] The IRT-estimated continuous S-shaped curve whose shape is defined by the A, B, and C parameters closely tracks the actual proportion correct for English field test kindergarten children (red triangles) and first graders (yellow triangles) across the ability range as well as the Spanish field test kindergarteners (gray squares). This item, which is representative of the majority of items in the SBRS, shows data and model curves with steep slopes ("a" parameters at or greater than 1.0) and measured ability levels increasing with increased grade level (i.e., children at higher grade levels with lower thetas have the same probability of a correct response as children at lower grade levels with higher thetas). Review of the classical item analysis results (r-biserials, distracters, etc.) confirmed that these items were all functioning well in the assessment.

Some SBRS items, however, did not show common functionality across languages. Figure 2 illustrates an example of this. Here the data from multiple samples administered the English version of the item are common-functioning and the model tracks the data well, while data from the Spanish version of the item (grey squares)

[1] The calibration focused on the kindergarten and first grade samples, and not the later grade levels, assuming the SBRS items would not be administered as a somewhat separate subtest beyond first grade.

[2] The following discussion of the results utilizes item characteristic curves that show the IRT results for some sample items. On the graphs, the horizontal axis corresponds to the estimated ability level, Theta, of students in the sample. The vertical axis, Probability, is the proportion of test takers answering correctly at each ability level. The colored triangles and squares show the *actual* percent correct (P+) for the field test participants in different grades, while the smooth S-shaped curve shows the *estimated* percent correct at each ability level derived from the IRT model. The A, B, and C parameters that define the shape of the curve (left-to-right location; steepness; asymptotes) are computed by the Parscale program as the best fit to the actual item response data.

appears to be easier for the Spanish-speaking group than for the English-speakers. Looking at the graph, at an ability level of -1.4, about half of the SBRS sample responded correctly to the item, where the English speakers of a higher ability level of -0.7 had the same probability of a correct response.

Of the 46 translated items administered, about two-thirds of the items were optimal for the SBRS, with high discrimination, a variety of difficulty levels, a range of content categories, and were common functioning with the English version. The remaining one third were not optimal for the SBRS for the following reasons: bad fit, not-common-functioning, too difficult, or low discrimination.

Recommendations

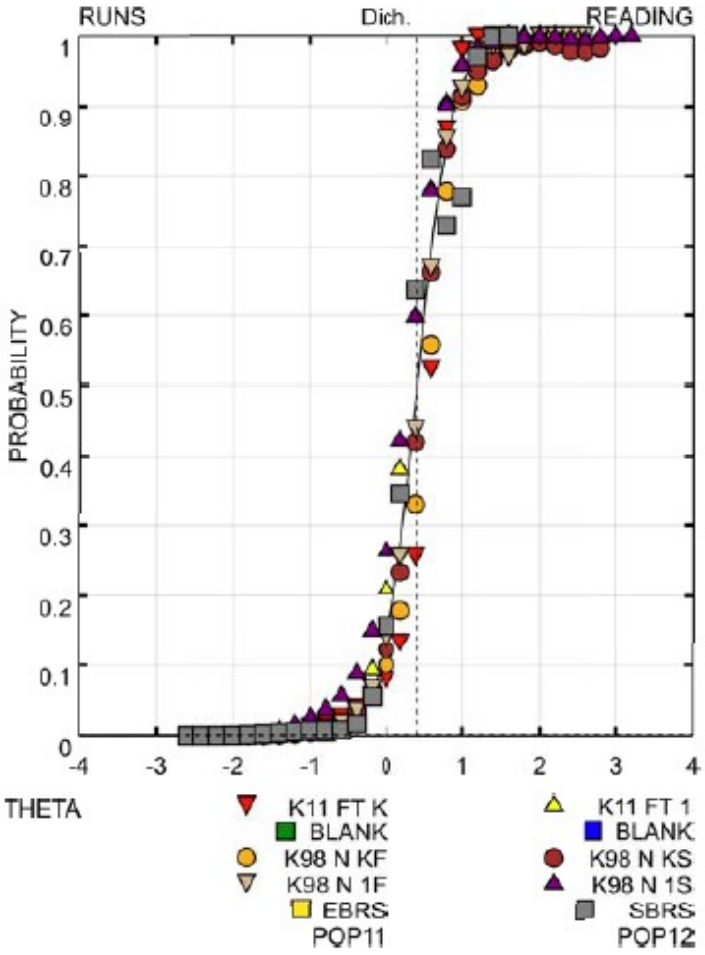
Based on the field test analysis findings, assessing Spanish-speaking ELL children using the SBRS items is recommended since the results indicated that an adequate number of items were common-functioning with the English versions, lending itself to develop an English score. Administering the SBRS in the national administration will also permit measurement of Spanish-speaking children's reading skills and knowledge in their native language, as well as comparison of Spanish speakers' basic skills in English and Spanish, when used in concert with the English Basic Reading Skills[3] assessment.

Next Steps

Upon receiving approval for these recommendations (or after subsequent revisions), we will proceed with selecting items for the SBRS. Delivery of the item selections will include a descriptive memo describing the analysis methods used, target ability and difficulty ranges, item overlap, and consistency with the framework.

[3] The English Basic Reading Skills (EBRS) is comprised of a set of items from the basic skills and vocabulary content categories that will be administered to all children in the national assessment, regardless of language spoken. More information on the EBRS is available in the memo on the EBRS of February 3, 2010.

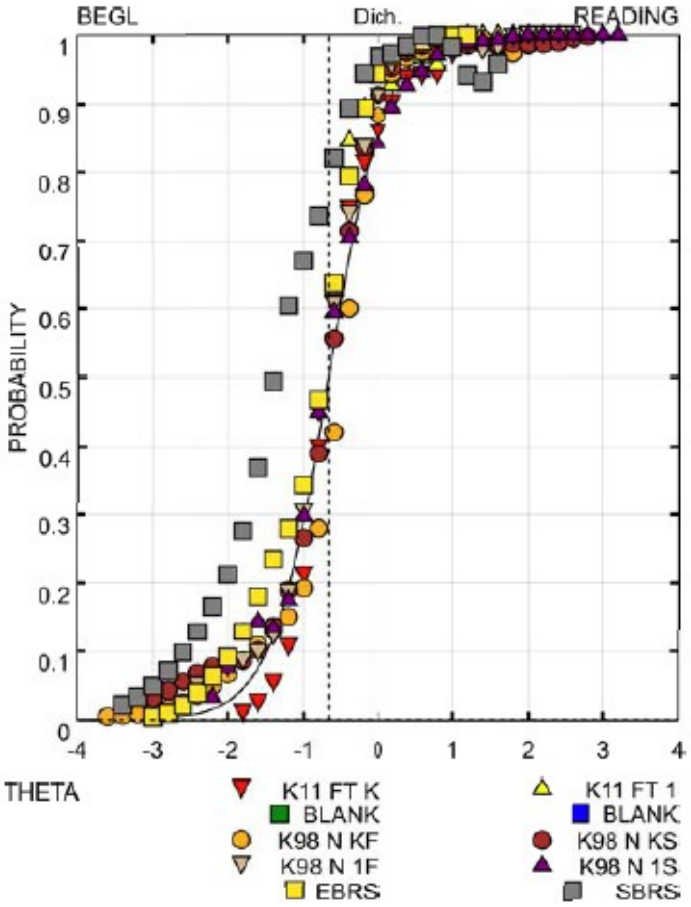
Figure 1: Item common functioning across languages



Parameters

	VALUE
A	2.6400
B	0.4206
C	0.0007
P+	0.4366

Figure 2: Item not common functioning across languages



Parameters

	VALUE
A	1.6002
B	-0.6722
C	0.0006
P+	0.7158