Workshop 1

Designing Assessments

Chair: Nicholas Zill
Presenters: Donald A. Rock, Phillip Fletcher, Paul A. McDermott, Jr.

Rock: The topic is how to develop a test to measure change, because often tests are used that were not developed specifically to measure change. Therefore, that goal has to be in mind from the start, otherwise the process is likely to fail. For example, one can get large negative correlations of initial status with gain. That kind of finding would occur due to ceiling effects when using a grade-level test, rather than a test designed specifically to measure change. Without an adaptive test to measure change, there will be biased estimates of change in children.

Two purposes for measuring change are to measure magnitude and area of improvement within a child's gains. Four important points must be considered: a) the measurement has to be an adaptive test, b) the test score scale must have multiple criterion reference points that ideally mark crucial developmental milestones, c) the test needs to be consistent with a particular model of development, and d) the test must cut across both age and grade levels.

Item response theory (IRT), scoring, and equating are also needed so that items' difficulties can be placed on the total test score scale. This kind of information can then be used to make an inference of where the child has made maximum gain. Every child will have a unique position on the scale, which is his locus of maximum gain.

Within an adaptive test, a child is presented with an item or set of items which are scored in real time, and based on this update of a child's ability level, a new item or set of items are selected to match the child's ability level. According to psychometric theory, more precise measurement can be made if the child's ability level and the difficulty of the item match. Giving items that are too easy or too difficult for a child tells nothing about the ability of that child, as he or she gets them all right, wrong, or makes guesses. Tailoring items to a child's ability level maximizes precision of measurement, increases reliability, and minimizes testing time. Without an adaptive test, the gain score analysis of classrooms populated by a large proportion of children at either end of the ability distribution in the fall or spring assessment would likely yield incorrect results of both student trajectories as well as relationships of gain with teacher, parent, or school process variables.

The Early Childhood Longitudinal Study (ECLS) used a two-stage computer-assisted adaptive test considered to be appropriate for the ages and abilities present in kindergarten and subsequent early grades. Why is there a need for multiple criterion reference points? Adaptive tests have the potential to estimate gain scores with virtually equal precision along the vertical test scores. They do not themselves provide information on which operational skills a child is making progress. When the adaptive item pool is developed, marker items to describe critical steps in the accumulation of language skills or mathematics can be identified and subsequently located on the various ascending points on the vertical test score scale.

The less the correlation between initial status and gain, the lower the correlation between process variables and gain. In order to have a complete picture of developmental change, one must not only measure how much a child gains, but also where on the vertical scale the child makes gains. The locus of gain can be made more policy-relevant if the vertical test score scale is behaviorally anchored at ascending points, with anchoring items reflecting milestones of learning complexity. In an adaptive testing situation, background and process variables tend to relate to where on the scale the change is occurring, rather than the amount of change. The ECLS battery is designed to be adaptive and provide reliable estimates of amount of change and also where the change occurs.

Fletcher: The focus of this presentation is on assembling short forms using an item bank. This technique is used in a number of Head Start research projects and also with ECLS Birth Cohort Study, which is for small children aged 9 to 24 months. How is IRT used to subset items for short forms? An item bank is used to do this. Precision is implied by small standard errors of reliability, that is, true score variance plus the error variance. If there is no true score variation in the population, then the precision of the test does not matter. These relative distinctions of rank are needed for meaningful analysis.

It is more realistic to step out of psychometrics and look at the other demands on testing. Shorter forms take less time with small children. The best content balance within tests is crucial. Each test should cover all of the content domains in the area. Broader subject matter coverage across different tests is key. All things considered, acceptable levels of precision and reliability for assessment purposes is needed. It is important to understand the group mean, that is, the growth over the course of the school year.

There are two major developments in the area optimal test design. One is about putting together an adequate item set, and the other is computer adaptive testing (CAT), which is a variant of tailoring the test to the difficulty level of the target population. Optimal test design is to select the most appropriate set of items to meet the overall test objectives when used with a specified target population. An adaptive testing strategy is often used, based on optimal sets of items. Many times it is acceptable to administer only the core item sets. One starts with the mid-range level of difficulty and if the child is in between the extremes of those items, then he or she is well measured. Therefore, about two thirds of the children only receive the core item set, about 15 % receive the basal items that are easier, and the ceiling items set is for another 16 % of children who perform well.

Usually with CAT, the test is stopped after relatively few items, resulting in less work for the child. In the National Reporting System, new developments are underway in the direction of computer adaptive testing. Today, IRT is the major paradigm in testing. The interesting thing about IRT is that the item now could be used as an interchangeable component in building a test. Therefore, an item could be added or subtracted, without changing the underlying scale metric, which is very important. Computers store all the item parameters on difficulty level, discrimination, and the guessing associated with each item. It can store content specifications so that there is a good balance in the content domain of the test. It can even store the actual text and graphics to be printed on forms.

Item banking is where item parameters are stored on a computer, and from the item pool, any number of alternate tests can be developed. The item banks also include a selection algorithm for automated test assembly. Any other test drawn from the same item pool produces comparable test results, reported in the same scale metric. Technical properties are usually known before the test goes into the field, so the precision and reliability of the test are known.

McDermott: The goal of Learning Express was to assess curriculum, focusing on literacy skills and mathematics. The problem with the norm-referenced tests is that they are fed by their markets. The best of these commercial tests tend to be driven largely by kindergarten and early elementary markets, so prekindergarten often gets the left over items. Even these norm-referenced tests are centered around the national 50^{th} percentile; unfortunately, children from challenged populations, such as those in Head Start, are likely to operate below the 50^{th} percentile. Therefore, the item pools that are available on these commercial tests become less relevant. A test whose item range was distinctly set for 3- to 5-year-olds in Head Start was developed that could be used for repeated measurement. The Learning Express was developed with four content domains: alphabet knowledge, vocabulary, listening comprehension, and mathematics.

The Learning Express is a direct assessment with a trained person doing the assessments. It is a skill-based, criterion measurement referenced to the new National Head Start indicators, as well as to the local state Head Start standards. Although not used as criteria, the items on the test are also mapped to other instruments, including the National Reporting System fall and spring item set. Each concept or skill being covered is aligned with other standards or norm-referenced tests that include the skill.

There are 4 measurement waves throughout the year. Initial assumptions concerning order of progressive difficulty were made on the basis of theory and on the nature of construction of some commercial tests. For the population studied, many assumptions were incorrect, and items had to be reordered. After administering the first item, if the child does well, he or she proceeds upward. With this form of adaptive testing, movement is up the scale until there are five consecutive failures, and then the test is stopped.

As with most IRT ventures, different models are assessed. Since the items are dichotomous, one, two, and three parameter models were reviewed. All of the scales now applied are two-parameter models that appreciate the relative difficulty of the items and the relative discrimination power of the items. It is not enough to look at group differences from one time and another. It is certainly not enough to look at differential growth at points in time. The test must be much more sensitive than that. For the Learning Express, internal reliability and concurrent and predictive validity were supported, and the test is sensitive to growth over brief intervals.