

Psychometric Foundations for Formative Assessment

By

Richard D. Schwarz

Robert Sykes

CTB/McGraw-Hill

**Paper Presented at the National Educational Computing Conference
New Orleans, June 2004**

Introduction

Before beginning a discussion of formative assessments, it is necessary to outline some different types of assessments and the nature of their intended inferences regarding student achievement. Summative assessments like those administered for the “No Child Left Behind” legislation are traditionally based on the annual collection of a relatively large number of data points (items) to provide a reliable test scores (and less reliable objective level scores). The intended inference for summative assessments concerns the change in achievement status for schools, districts, or entire states. By contrast, formative testing occurs frequently (weekly, bi-weekly, monthly) for a relatively small number of data points (items). This type of assessment is intended to provide teachers with regular and detailed information that characterizes student proficiency and supports clear instructional prescriptions within a classroom. Formative assessment can simply be defined as the provision of better measurement in the classroom that can positively impact instruction and learning. Formative assessments must match district/state standards and have technical rigor (content & psychometrics) in order to meet federal requirements and standards for educational and psychological testing. The purpose of formative testing is to supply detailed information:

- (a) about what students know and do not know at a highly granular level,
- (b) to support prescriptive teaching information at the student, class, and school level,
- (c) closely aligned to state standards,
- (d) to support instructional effectiveness, and
- (e) that indicates how well students are prepared to meet the demands of state standards and annual state summative assessments.

Diagnostic assessment is the identification of the skills, concepts, or cognitive attributes that a student might possess that can be used to inform the next episode of instruction.

The first step in the construction of a diagnostic assessment is how to represent

knowledge in a domain (e.g., algebraic relations). A second step is to articulate the states or levels of understanding held by individuals in this domain. Often a cognitive model is used to define this problem space and to specify various approaches for solving the problem. However, cognitive models typically only permits inferences about student learning based on task performance within a single well-defined domain (e.g., spatial rotation). Therefore, a fine level of item specification is necessary for diagnostic assessment compared with other types of achievement tests.

Finally, the most sophisticated types of assessment are intelligent tutoring systems. Intelligent tutoring systems include assessments designed to diagnosis a student's relative strengths and weaknesses, prescriptions for the next episode of learning, and practice. These activities are all targeted at the student's current level of understanding and are components embedded within the system. This type of system is predicated on the assumption of relatively long periods of interaction between the student and the tutoring system.

Formative assessment fills the middle ground between the granularity of inferences between summative and diagnostic assessment that best supports instructional decision-making in this context. The remainder of this document will focus primarily on the features of formative assessments necessary to provide a sound psychometric foundation.

In constructing such a system, there are many potential options that could be adopted and this system can be flexible deployed in a number of different configurations. Some fundamental design philosophies for a formative system are:

- high technical rigor (validity and reliability),
- assessments that have depth and breath which include constructed-response items,
- assessments that can be given at multiple times in the school year that can demonstrate student growth,
- both paper & pencil and on-line modes of assessment being offered,
- fine grain alignment to content standards,

- reporting information that is instructionally relevant that are powerful but intuitive to users,
- immediate reporting of test results
- incorporates elements of universal design,
- the use of local educators in developing and customizing some aspects of the system,
- accrual of longitudinal information to guide decision-making
- a system that continuously evolves and improves over time.

As this system has been described, it clearly entails both an on-line test administration and reporting system. Previous efforts at diagnostic assessment placed too much burden on the teacher to administer, score and interpret assessment data. A well-conceptualized assessment system can avoid the pratfalls of these earlier attempts by utilizing the convenience afforded by computers linked to the Internet where many features can be automated and updated periodically. It is also recognized that not all districts or schools will have the technology infrastructure to support such as system. Therefore as a bridge to this emerging technology, schools need the capability to flexibly administer tests in traditional paper and pencil mode and have the ability to scan these results into the assessment information system and thereby derive all the system benefits.

Some Technical Characteristics – or the lack there of. It is important that information provided by formative system be validated. The ability of formative assessment system to adversely affect learning by making implicit or explicit instructional prescriptions not supported by data should not be underestimated. Validation for a formative assessment includes an examination of the items that comprise the assessment and an alignment of that test content to standards, obtaining item and test statistical characteristics and may even involve steps such as ascertaining the usability of score reports. All the traditional sorts of reliability and validity information will be obtained for these tests consistent with the Standards for Educational and Psychological Testing (1999). Formative tests should have technical rigor (validity and reliability) normally associated with summative tests (high item discriminations, appropriate difficulty, DIF minimization, etc). Currently,

many formative tests are completely deficient in supplying any reliability and validity information.

The use of a scale can facilitate inferences with respect to student growth. A scale can be viewed as a developmental continuum where as students develop new capabilities, they move along the continuum, as demonstrated by their scale score. IRT scale scores are units of a single equal-interval scale where higher scores indicate greater proficiency in a domain. This equal-interval property makes scale scores especially appropriate for various statistical purposes such as averaging that permit direct comparisons of change for students classes, or schools. If scale scores are increasing over time, then student achievement is assumed to be improving. It is possible to make inferences with respect to growth of knowledge in a domain without a scale (see Doignon & Falgagne, 1999).

A Test Design Using Two Types of Integrated Formative Tests

In order to get some understanding of what might be entailed in a formative assessment program, two types of tests are described here. This example is given primarily to further delineate the characteristics of formative test with the recognition that many other potential test configurations are possible.

Benchmark Tests. The essential idea for formative assessment is that it is directly integrated with instruction rather than a unity event that takes place annually. Since testing is much more continuous in a formative system, it is important to have a test design to that allows student growth to be accrued and evaluated periodically but does not result in undue amounts of testing time. To meet a variety of different purposes, two types of integrated tests are described in this formative design called Benchmarks and Growth Strand tests. Progress at the district or school level can be monitored with Benchmark tests. Benchmark assessments can be administered as pre- and post tests at designated times to determine what students do and do not know against state standards and how much growth has been demonstrated in the school/district. Each Benchmark test can contain enough items to be reliable and to articulate to the scope and sequence of

instruction that are somewhat graded in difficulty that match a given state’s standards. The Benchmark tests survey could survey a state’s standards in the same way that summative assessment does. Having Benchmark tests that more closely match the scope and sequence of instruction that match to a state’s content standards can provide information that can more effectively impacts instruction and learning. Using a pre-post design, two formative Benchmark Tests can be created that are shown in Figure 1. The yellow box denotes the summative test and the green boxes indicate the Benchmark formative tests. These tests are designed to be given in a pre-post configuration with the first Benchmark given early in the school year followed by one that is in close proximity to the summative assessment. Figure 1 shows arrows from the Benchmark tests to the summative assessment that indicates some type of linkage. The line between the two Benchmark tests shows that these two tests could be on the same scale after the completion of a calibration/equating study. The term “Item Bank” in the figure refers to an item bank being made available for those teachers that wish which to construct their own tests for classroom use. In addition to monitoring progress at the district, school or class level, the Benchmark tests can also be used to determine students’ respective strengths and weaknesses that can be further targeted through the use of Growth Strand (Level II) assessments at the classroom level.

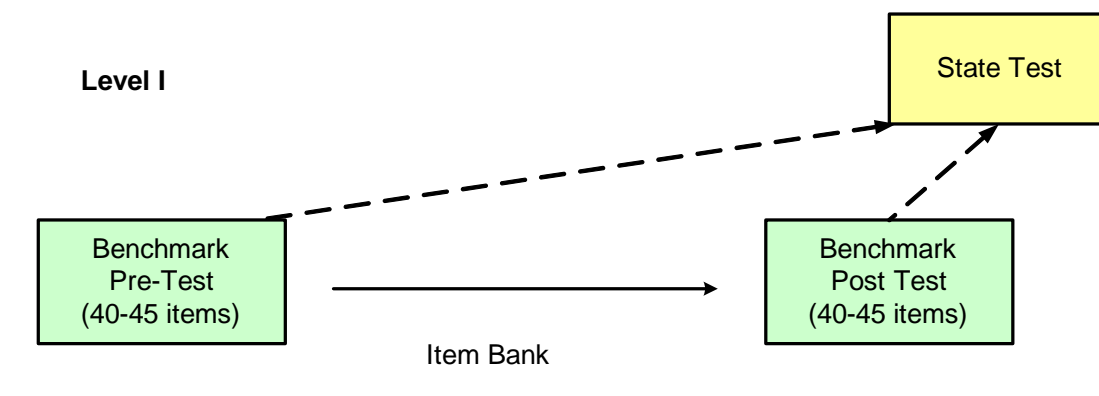


Figure 2.

Articulation of Two Benchmark Tests and the Summative Assessment

Growth Strands. The Growth Strand tests help form a truly formative system that can inform classroom instruction and allows students to monitor their own growth. The intent of these Growth Strand tests is to give teachers “diagnostic” information that reflect a student’s understanding for a given content standard that directly relates to instructional interventions. Figure 2 uses some NCME Mathematics Standards as an example to show how the Benchmark tests articulate with the Growth Strands Tests, which are defined by a more granular level in the content hierarchy. Four different Growth Strand tests represented by Numeration, Measurement, Estimation and Functions are given as examples.

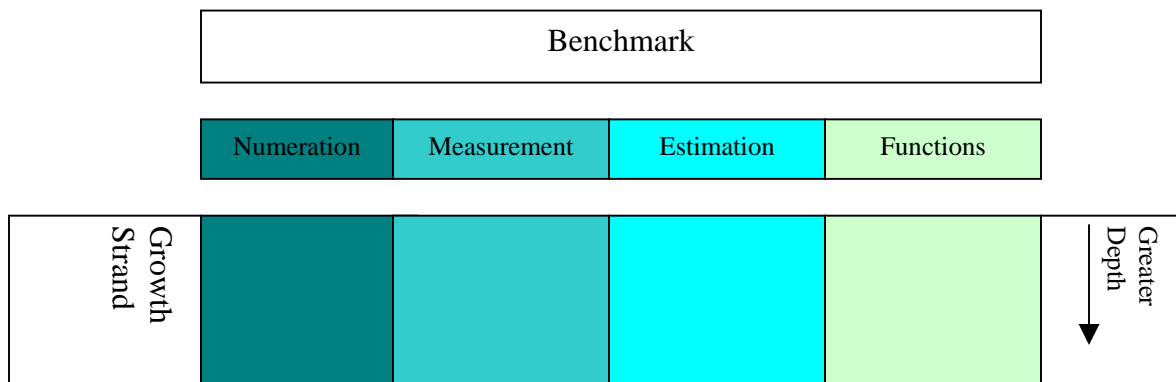


Figure 2.
Articulation of Benchmark and Growth Strand Tests

Any given Benchmark test can only sample the content for a standard horizontally at a certain level of coverage (say 4-6 items) based on a test with 40-45 score points, shown in the bar above. Each Growth Strand test can contain approximately a smaller of items in a single “testlet” that permit greater coverage in a content domain. Several hundred items in each Strand will be available in total that permits greater depth and breath of domain coverage with a range of item difficulty, see Figure 3. Strands tests supply more “diagnostic” information than Benchmark Tests, which in turn provide more granular, subskill information than the summative assessment since they better reflect the scope

and sequence of instruction. Growth Strand tests are designed to give teachers more granular information since content is vertically sampled in single standard that directly relates to classroom instruction. The Growth Strands should match any state's standards due to their greater level of granularity. However, naming conventions, content hierarchy, and scope and sequence of instruction will differ among state content standards. If provisions need to be made for larger instruction units or review then it is possible to aggregate two or more levels of the Growth Strand tests together. Growth Strand Tests are given on a "need" basis. One use case might involve a teacher giving a Growth Strand Test when a class begins a new unit of instruction and again upon completion of the unit in a pre- and post- test design. Growth Strands can be administered to students who are working at their pace/level in order to target their ability and individualize subsequent instruction. The Growth Strand tests can replace the classroom types of teacher tests with ones that can better characterize student learning that can be used along with other types of information from the system such as instructional interventions. The scenarios described here will need to be further developed in conjunction with teachers. There is considerable flexibility in how the Benchmark and Growth Strand Tests can be configured in this on-line system. For instance, one school system may only elect to administer the Growth Strand tests. Another scenario is to have the Benchmark tests from the previous year's spring test given again very early in the school year as a review test. Finally, a very important consideration is a discussion of how the Benchmark and Growth Strands tests are used in conjunction with information from the summative assessment (see section linkage to the summative assessment and strategies for augmenting summative assessments).

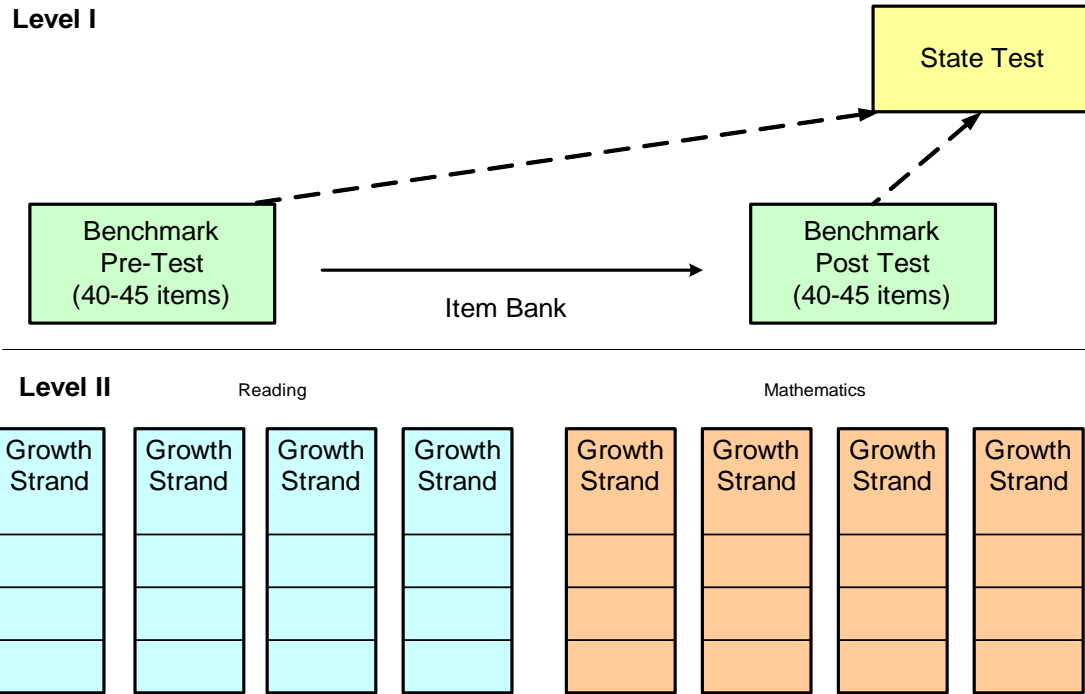


Figure 3.

A test design showing Growth Strand tests for Reading and Mathematics that articulates with the Benchmark and Summative (State) Assessment.

Adaptive Testing Systems

Item adaptive tests are well known and have been used operationally for a number of years primarily in high-stakes licensure examinations. In item adaptive testing, a student’s test performance determines which items or blocks of items (stage adaptive) are administered such that ability can more precisely be targeted with fewer items administered. An item-adaptive system could be a useful tool for formative assessments since some student’s placement can be well or behind of the classroom and instructional differentiated instructional interventions need to be designed accordingly. That is, if a student’s position can be more precisely defined on the growth strand then appropriate instructional interventions can be planned. Other types of stage adaptive tests could be constructed for paper-based administrations but this would be more unwieldy to administer. Item adaptive administrations could also be implemented for Benchmark tests.

Item Types

Since a mixture of selected (SR) - and constructed response items (CR) will likely be present on summative tests, formative tests need to contain both item formats in order to sample the content standards in a comparable fashion. At this point, there is not technology available in which all the varied types of constructed-response items present on paper- and pencil tests can be scored automatically and results returned instantly and economically consistent with classroom needs. The constructed-response items for formative assessments need to be ones that can be adapted to an on-line environment. Currently, two types of constructed-response items can be administered on-line. Ones that are text based such as writing prompts which can be scored using algorithms such as latent semantic analysis (Landauer, Laham & Foltz, 2003). Other types of CR items in which the range of student responses are finite, which can be scored automatically using expert judgment, called constrained constructed-response. These types of items formats could consist of various types of “drag & drops” or “fill-in-the-blank”. These CR formats enable some standards to be covered that could only be indirectly assessed with only selected-response items. Currently, there is no automated scoring processes that can duplicate the full range of CR items formats administered in paper- and pencil modes.

It is conjectured that other types of innovative items types that use audio, video, and other types of media will be eventually be available for use in formative assessments that can more realistically reflect tasks. The use of simple four-choice, multiple-choice items under utilizes the potential functionality of the computer. For instance, it is easy to conceptualize items in which a student has to construct a bar graph for a particular data set using the computer using “drag & drop” functions in the interface. Other types of items could have students view a brief video, and then answer some multiple-choice items and write a response to a writing prompt. Incorporating these types of items will further enhance the construct validity of the tests. Of course, it is likely that these types of items will not be represented on the summative test at least for the foreseeable future.

Linkage to the Summative Assessment

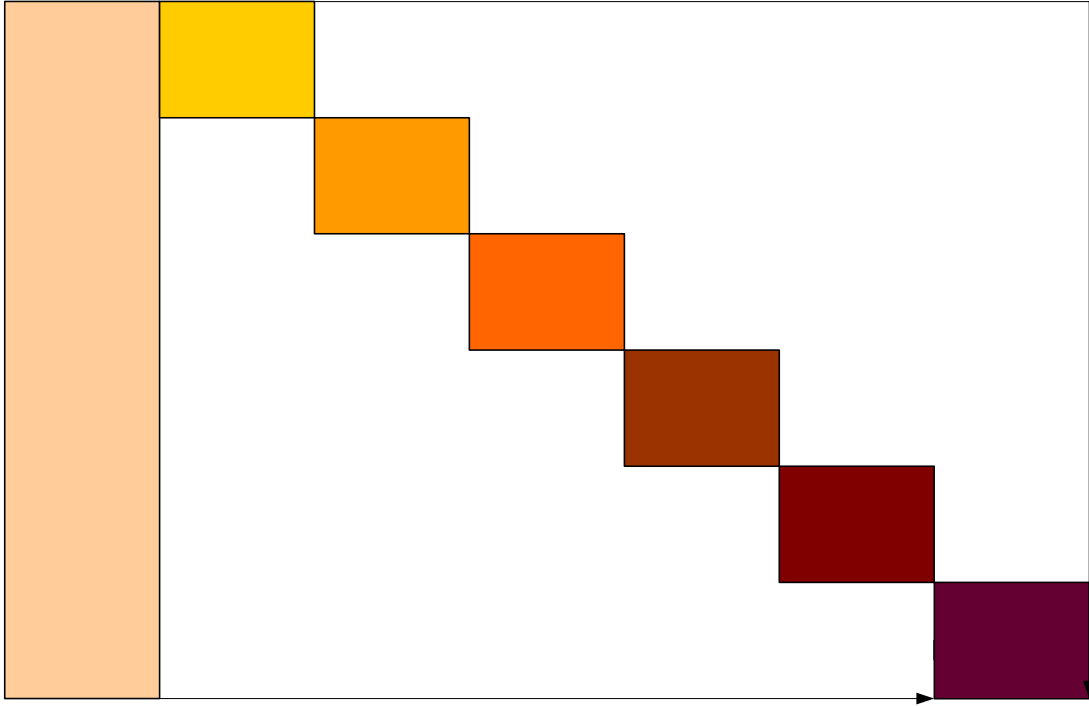
The most critical question that a formative assessment must answer given the requirements of NCLB is whether students are on target to be Proficient on the summative assessment. Users of formative assessments need to interpret performance information in light of this requirement. Traditionally, tests that are on different scales that need to be compared are equated or linked using statistical methods such as equipercentile equating. There are many limitations to this type of approach for linking formative and summative assessments. A more useful approach might be to let users have access to the data from summative and the formative tests in an interactive database and allow them to make hypotheses about their student using structured queries, develop hypotheses and collect additional evidence over time. Having a single years worth of data for summative and formative tests, one could ask “How high do the scores on the formative assessment have to be in order to be confident that students will be proficient on the summative assessment in either math or reading?” A first step might be to examine the range formative performance of students who were Proficient in either Math or Reading. This is called a concordance study. One approach might be to take students who were Proficient in Math on the summative assessment and then, say, compute their median score on the formative assessment. If students obtain the median score on the formative then one could be somewhat confident that they will be Proficient on the summative assessment. This allows users such as principals or teachers to make queries about their students where they have knowledge about their capabilities. This is an evidence-based approach in which information can be gathered across administrations in order to make stronger inferences based on multiple observations. When multiple observations are available for each student, longitudinal models can be used to “predict” performance. In addition, there are many psychometric models used within the context of longitudinal data (Singer & Willett, 2003) that can be utilized to examine growth expectations from formative assessments.

Psychometric Models for Formative Assessments

There has been increased interest in profiling the characteristics of students taking by positing different strategies for solving problems and to identify the types of solution approaches used by students. Since the intent of formative tests is to build a more diagnostic/prescriptive assessment system, a number of psychometric models have proposed in the literature that can be used to address this need. These models provide a richer definition of a student's knowledge, concepts, and skills than achievement tests which are more focused on the rank ordering of students. Some of the better-known psychometric models have been formulated under the rubric of diagnostic assessment are very briefly summarized below. Most of these models can be characterized as having more model parameters than is the case with traditional IRT models. These models are roughly ordered by the number of model parameters specified.

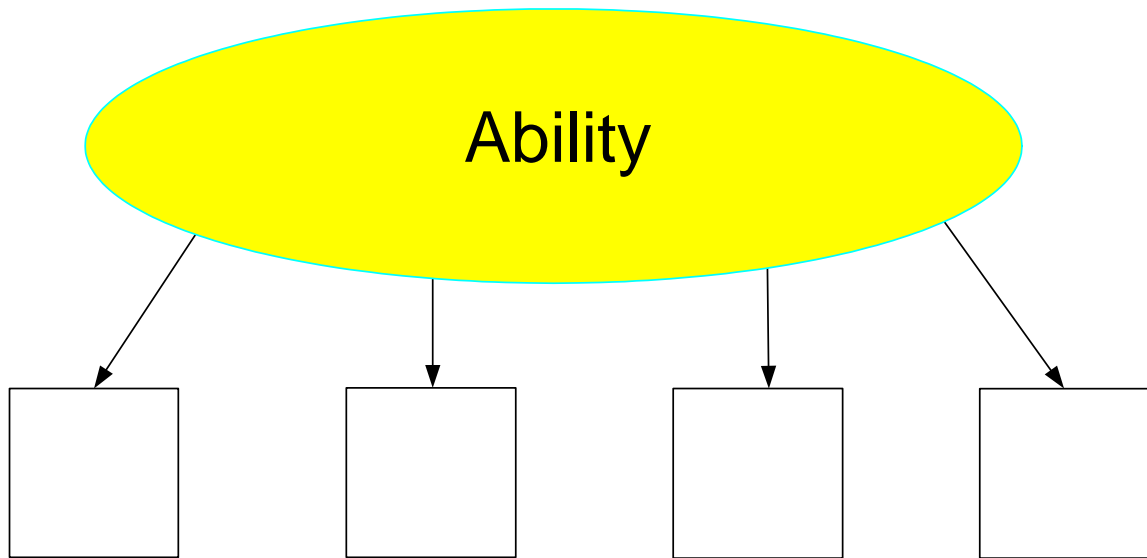
Enhanced Classical Test Theory. This model uses raw scores but with the provision of matrixed content. This allows users to gauge proficiency efficiently across an array of curricula in a subdomain (e.g., a standard) while facilitating inferences about what students “know and can do”. This avoids unwarranted assertions about group performance based on a small set of items. This model could be used in conjunction with those listed below.

Enhanced Classical Models



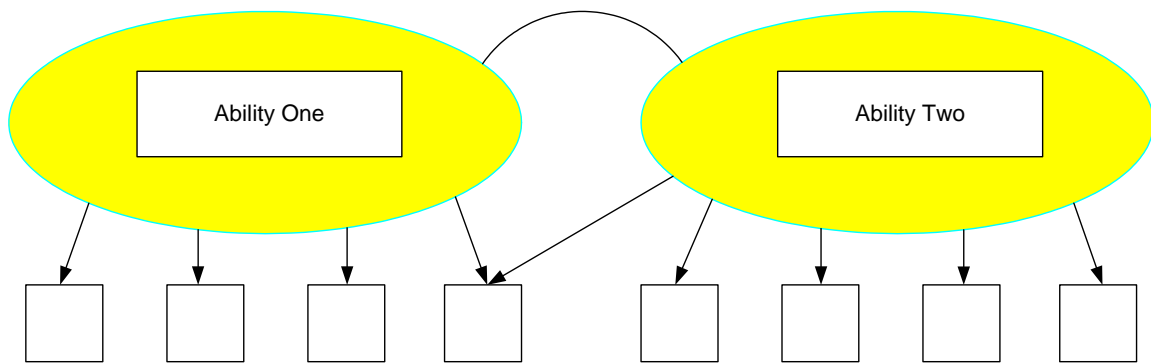
Unidimensional IRT. These are the same models and score types that have been extensively by many types of assessments. A variety of familiar score types and interpretations regarding student growth can be made for formative assessments based on this model.

Unidimensional Models (IRT)



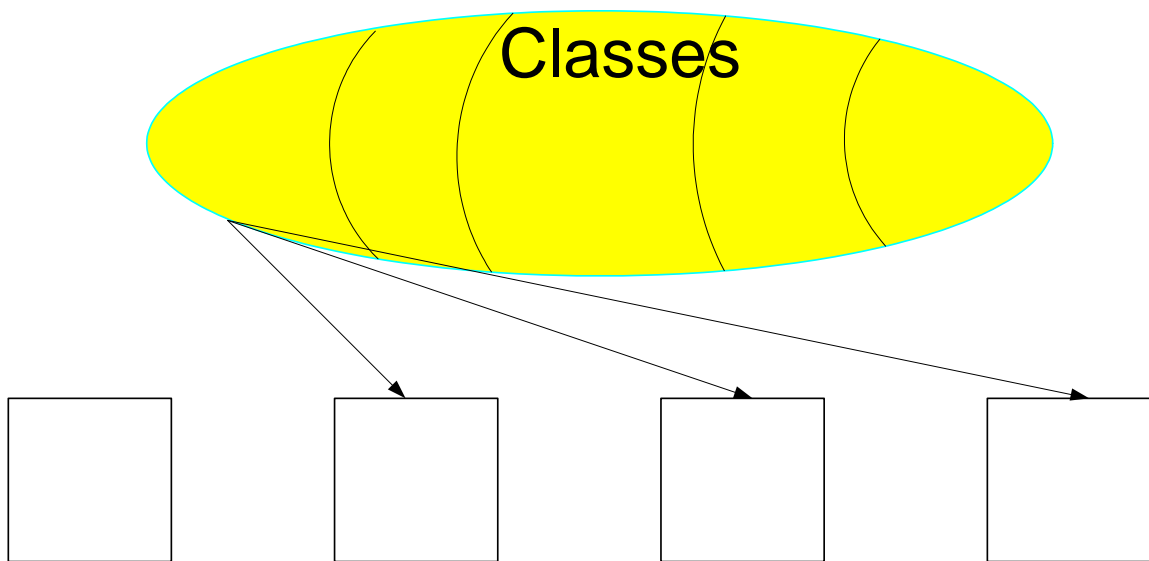
Multidimensional IRT. These models are within the IRT family but postulate two or more abilities that account for performance. For instance, some story problems math items require the item to be read and a solution computed. These abilities may correspond to groups of items with similar content or they may relate to characteristics of students such as the strategy employed to solve problems. Each student receives a score on each dimension (ability).

Multidimensional Models (IRT)



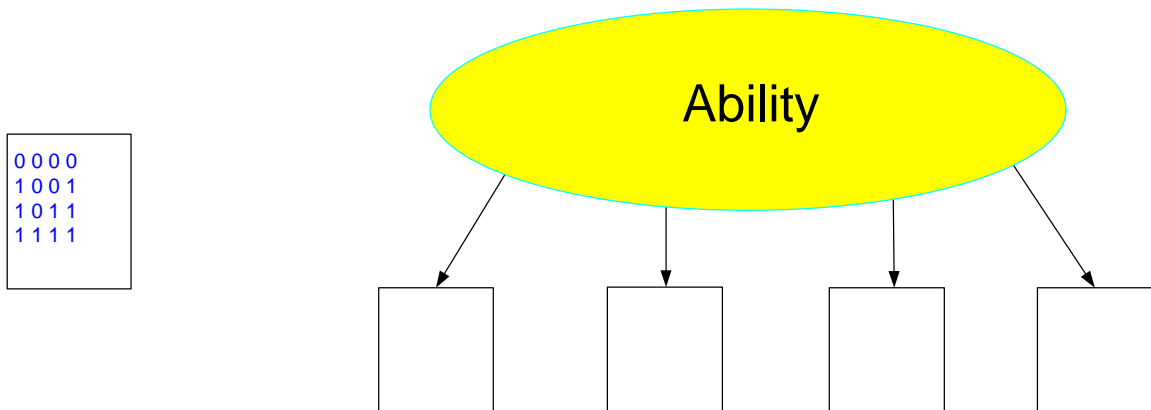
Latent Class Models. Ability is considered to be a discrete variable rather than a continuous one. A latent class of masters and non-masters could be formulated to account for test performance, for example. Other types of models combine latent class with IRT called mixture models.

Latent Class Models



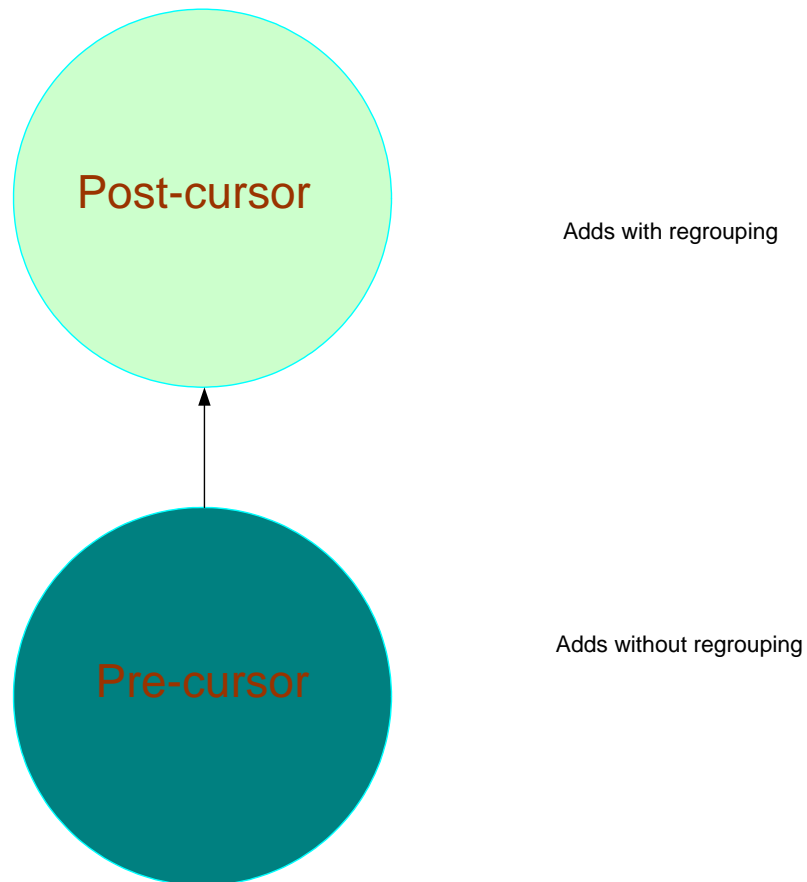
Unified (Fusion) Model. This is also a type of mixture model with a discrete latent class attributes along with a continuous ability (i.e., IRT) estimate. The intent of this model is to use the respective strengths of a theory driven component, which are discrete/deterministic with a stochastic element (IRT). The model is able to predict individual attributes of students (e.g., the student knows the rule).

Fusion Model



Bayesian Belief Networks. (Adaptive Probability Networks). These models use Bayesian inference to establish a network of relationships between items or sets of related items. This model can be used to empirically validate precursor and postcursor concepts (smallest teachable concepts) that are hypothesized by the researchers or content expert. A student can be located in this network and probabilistic inferences can be made whether he/she knows a targeted concept.

Bayesian Inference Models



Strategies for augmenting or replacing summative assessments with formative ones

The formative system that has been described here affords a rich amount of assessment information than is present in a summative test by design. Formative tests offer multiple observations as opposed to the single one based on the summative assessment. Studies could be undertaken to determine how well the formative tests “predict” the summative assessment. These investigations could be used in a longitudinal framework using Benchmark or Growth Strands as the prior information. Multiple formative tests also afford much better coverage of a state’s content standards. There could be an opportunity to supplement or replace a summative assessment with a formative one such as the one described here if the linkage is sufficiently strong. There would be likely two requirements for doing this from a federal perspective. First, there has to be a match to state standards and secondly there must be technical rigor (i.e., sufficient reliability and validity). There should be sufficient evidence and technical rigor for in the formative assessment system described here. One strategy might be to weight the information from the formative and summative assessment into a composite score. The summative and formative score could be given arbitrary weights of ½ each. There are also ways to statistically weight scores in a composite. This approach has the advantage of not having to give all the weight to a single summative score. A more far-reaching option would be to completely replace the summative with the formative assessments. At the least, if any substantive changes were made to the NCLB legislation in the future, users would have a viable alternate program already in place.

Professional Development

A part of professional development, teachers could be partnered with to write items, design student profiles and instructional interventions. The assessments described here will require a large number of items to be written and validated. As a professional development activity, teachers could be trained and certified by to write items for this

assessment. The result will be items that closely match the intent of the state/district standards and a fuller understanding of inferences related to test scores.

Some additional steps in the test construction process would be to collaborate with teachers by developing links to practice/instruction and developing meaningful instructional interventions. One method might be to develop various profiles of students with different abilities and then use teachers to develop the appropriate instructional intervention. Psychometric models can be used to construct student-learning profiles. Statistical profiles identify particular types of student problem solving strategies and knowledge states. Researchers could collaborate with teachers using this information to delineate various sorts of instructional strategies. This entails a process that iteratively improves over time. Teachers could select from instructional strategies and practice lessons, which are ordered in terms of usage from high to low like Internet search engines. Using the described assessment system could allow students to monitor their own growth with prescribed interventions.

Reporting

As a principle formative assessments will be facilitated by a focus on test content and observable proficiency on content (e.g., performance level descriptors, domain scores, raw scores) and in turn make statistical scores more invisible. Teachers, students, and parents need intuitive information that is less statistically oriented and is focused more on content and learning. In principal, the report information should be powerful but intuitive to users. The reporting system should be able to recommend instructional strategies for teachers or to help enable students to monitor their own growth.

When test items are scaled, then more robust depictions of student achievement are supported that reference general abilities required for success on items above and below the student's scale location. Since multidimensional models support more systematic and detailed analyses of components of proficiency, such performance level reporting could become more detailed, more useful, and more prescriptive. This observation applies to a broad class of "observably grounded" score types, including domain scores and objective

scores. As described previously, teachers and student need information for instructional interventions.

Formative like summative assessment has a number of additional reporting needs. First, there is still the need to report information for various levels in the hierarchy such as individuals or entire classrooms. For schools and districts, it is also necessary to view change for various aggregations. In the system that has been described, longitudinal information is a prominent feature since testing occurs with much greater frequency. Therefore, the system must have the capacity to store and display longitudinal information. As described previously, the reporting system should have the ability to compare performance from the formative and summative assessments.

Conclusion.

Educational measurement has primarily focused on summative assessments that have become a central outcome measure for school systems. Teachers need good and timely information that can be used to positively effect instruction and student learning. The opportunity to make major contributions to the quality of formative assessment used by teachers to benefit students is great. On the other hand, the opportunity to make major contributions to the quality of formative assessment as individual teachers regularly practice it is great, and taking advantage of this opportunity does not necessarily require complex technical approaches or sophisticated psychometric models. When the features described here are thought of holistically, a revolutionary change in assessment is possible.

References

Doignon, J.P. & Falmagne J.C. (1999). *Knowledge Spaces*. Springer:NY;NY.

Landauer, T.K., Foltz, P.W. & Laham, D. (2003). Automatic Essay Assessment. *Assessment in Education*, 10, 295-308.