Building a Large-Scale Educational Assessment System

Principles and Procedures

Goal of this Presentation

- To provide an overview of the design, development, and implementation of large-scale educational assessments

Steps in Building an Assessment System

- Determine the purposes of the tests and intended uses of the scores
- Specify the test content and target populations
- Develop test items
- Perform pilot and field testing of items
- Conduct item analyses
- Assemble tests

Steps in Building an Assessment System

- Administer tests
- Perform further psychometric analyses
- Set achievement levels if desired
- Score tests, create reporting scales, and equate across forms or grades
- Gather reliability and validity evidence related to intended uses of scores

Steps in Building an Assessment System

- Develop and distribute score reports
- Document procedures and results in a technical report

Purposes and Intended Uses of Test Scores

- There are a number of possible purposes of a statewide assessment system
- The purpose of the assessment must be clearly stated before appropriate tests can be developed
- Purposes vary by state and are determined by policy-makers and stakeholders

Purposes and Intended Uses of Test Scores

- Stated purposes in technical reports of various states:
  - provide data on student achievement to meet federal and state accountability requirements
  - provide information regarding student and school performance to parents and the public
  - provide information to support curriculum evaluation and improvement
  - provide information about equitable educational achievement across subgroups
  - monitor student growth over time (CT)

Purposes and Intended Uses of Test Scores

- Stated purposes in technical reports of various states:
  - set high expectations and standards for student achievement (CT)
  - provide measures of student achievement that will lead to improvements in student outcomes
  - identify students in need of intervention
  - help determine competency for the awarding of high school diplomas (MA)

Purposes and Intended Uses of Test Scores

- Stated purposes in technical reports of various states:
  - Assist in the identification of educational needs at the state, district and school levels (MS)
  - Assess how well districts and schools are meeting state goals and minimum performance standards (MS)
  - Provide a basis for comparisons among districts and between districts, the state and the nation (MS)
  - Provide data that can be used to aid in the identification of exceptional educational programs or processes (MS)

Specification of Test Content

- The test content must align with the State's grade-level academic standards in terms of content and cognitive complexity
- Claims about what students know and can do that will be made on the basis of test scores must be clearly stated
- A test blueprint must be developed that describes in detail the structure of the assessment with respect to coverage of the content standards and claims

Item Development

- Item and task specifications must be written to produce items that will provide evidence-based support for claim scores
- Item specifications describe the types of evidence that should be obtained for each claim
- These specifications provide models for writing items by describing the knowledge, skills, and processes to be measured by item types aligned to particular claims

Item Development

- Content and measurement experts work together to develop specific items according to the specifications
- SBAC used hundreds of practicing teachers of ELA and Math to write items for the assessments
- Teachers were trained with respect to the content specifications, item and task specifications and item-writing techniques, and received feedback from professional item writers
- Item-writers had to submit a sample of items of adequate quality to be certified to continue with item development

Item Development

- High-quality item development is a complex, time-consuming, costly procedure with many quality control checks
- Production of data displays, graphics, artwork, or other visuals and integration with the item text requires meticulous attention

Item Development

- Once items are written, they undergo review prior to administration
- Items are reviewed for accessibility, sensitivity, and content
- Accessibility reviews focuses on aspects of the item that may negatively affect a student's ability to demonstrate their knowledge (e.g., unnecessary complexity in text and visuals, poor organization and/or item layout)

Item Development

- Sensitivity reviews focus on content that may negatively affect a student's ability to answer the item because of their background (e.g., stereotypical portrayal of ethnic or cultural groups)
- Content reviews focus on alignment of stimuli, items, and tasks to the content specifications and required depths of knowledge, along with checks of accuracy of content, answer keys, and scoring materials.
- Items flagged for any of these concerns are either revised or removed from the item pool at this stage

Pilot Testing

- Pilot testing is typically administration of subsets of items to small samples of students to identify issues with items such as inappropriate difficulty, lack of discrimination, or other unanticipated problems
- SBAC used pilot testing to try out new item types and assess any problems students had in responding to them
- The SBAC pilot test was on a large enough scale to provide data that informed the choice of psychometric model to be used for the full item bank

Field Testing

- Field testing of items is conducted on a larger scale to determine whether the items are functioning as intended and to obtain information on their psychometric properties
- In existing testing programs, field test items are embedded within the operational test so that students respond to the items in the same way as they do to operational items

Field Testing

- Each student takes a few field test items, which are not counted towards the student's score
- Typically, 1000-2000 responses per item are obtained to provide reliable information about the item's characteristics

Field Testing

- For new testing programs where there is no existing operational test, stand-alone field testing is performed
- In this case, students take a test comprised solely of field test items
- Different students take different subsets of items
- Students know that their score on this test does not count, hence lack of motivation may affect performance

Field Testing

- SBAC field-tested over 15,000 items across grades and subject areas to about 1.7 million students
- Each student took a “linear-on-the-fly” computer-administered test (LOFT) comprised of about 50 items assembled to meet the test blueprint with respect to content
- Items were randomly chosen sequentially from the total pool with the constraint that the final test had to meet the content blueprint
- This design allowed all items to be administered to about 1200 students

Item Analysis and Model Selection

- The field test data is critical to the construction of the final test or item bank
- Items that appear to be measuring something other than the intended construct are identified and removed
- Analyses are performed to identify any items that show differential functioning across gender or race groups or other identified subgroups
- These items are removed from the item pool

Item Analysis and Model Selection

- One or more psychometric models are fitted to the data, and a final model is chosen
- These psychometric models assume that there is a statistical relationship between a student's proficiency and their performance on each test item (e.g., a student with this level of proficiency has an 80% chance of answering this item correctly)

Item Analysis and Model Selection

- The student's probability of answering an item correctly depends on both their proficiency and the characteristics of the item
- The psychometric model includes parameters that represent characteristics of the test item, such as its difficulty and discrimination (how well it separates those who know from those who don't)
- Estimates of the item parameters are obtained based on the field test data and these are then used to construct operational forms of the test

Item Analysis and Model Selection

- Estimates of the students' proficiencies can also be obtained, but they are often not reported because the purpose of the field test is primarily to obtain information about the test items
- When the items are administered operationally, their difficulty and discrimination will be used to estimate the proficiency of the student
- Students who answer difficult items correctly will obtain a higher proficiency estimate than students who answer those items incorrectly, and highly discriminating items will carry more weight than poorly discriminating items

Test Assembly

- Once all items have been “calibrated”, final forms of the test can be assembled
- The test blueprint provides detailed specifications with respect to the proportion of items in each content area, at each level of cognitive complexity, of each item type, response format, and type of scoring that must be used on the test

Test Assembly

- Knowing the difficulty and discrimination of the items allows test developers to choose a set of items with good discrimination and that are of appropriate difficulty for a particular group of students or for a particular purpose
- Multiple forms of the test can be constructed to have similar characteristics

Test Assembly

- The advantage of the psychometric models is that once all the items in the bank have been calibrated and their parameters are on the same scale, it is possible to compare the proficiency scores of students even if they have taken different forms of the test
- This feature forms the basis of adaptive testing, where different students take different items with difficulty levels matched to the student's proficiency

Test Administration

- Test administration procedures must be specified to ensure consistent and standardized administration of the test
- These procedures are documented in a test administration manual
- Standardized administration ensures that no irrelevant factors related to administration affect students' scores
- Inconsistent administration procedures reduce the validity and reliability of test scores
- Procedures must be specified to ensure test security and protect the integrity and confidentiality of the test data

Test Administration

- Test administration manuals also specify permissible accommodations
- Training materials should be provided to ensure that test coordinators and administrators are prepared to properly administer the assessments
- Test administrators must also be trained in the use of the technology used to deliver computer-based tests
- Contingency plans must be specified to deal with testing irregularities

Post-Administration Analyses

- After the operational administration, item analyses are carried out again to identify any problematic items that should be removed from the pool for future purposes due to poor functioning or differential functioning across subgroups
- The psychometric model may be refitted to obtain updated, more accurate estimates of the item parameters before final proficiency scores are computed

Achievement Level Setting

- Either based on the field test data or the operational data, achievement levels are defined and cut-scores are determined for classifying student into achievement levels
- Achievement level descriptors are defined through consensus by panels of content experts (teachers, educational administrators); these are detailed descriptions of what students at a given level should know and be able to do

Achievement Level Setting

- Setting achievement levels involves convening panels of stakeholders (teachers, school administrators, higher education faculty, business and community leaders, parents) to review the test items and reach consensus as to the proficiency score that distinguishes students at one level from students at another
- SBAC used both online and in-person panels
- Over 2600 people participated in online panels

Achievement Level Setting

- There are several methods for setting achievement levels; the most widely used method is called the Bookmark method
- Using the Bookmark method, the items on the test or a representative set of items are arranged in a booklet ordered by their difficulty from easiest to hardest

Achievement Level Setting

- Panelists are asked to work their way through the booklet in order and for each item, answer for themselves questions like this:

  1. What skills must a student have in order to know the correct answer?
  2. What makes this item more difficult than preceding items?
  3. Would a student right at the threshold of this achievement level have at least a 50% chance of earning this point?

Achievement Level Setting

- If the answer to the third question is Yes, the panelist continues to the next item
- If the answer to the third question is No, the panelist places a bookmark at that item to indicate that they would not expect a student at the threshold to be able to answer any item of this difficulty or greater difficulty
- The difficulty of the item indicates the proficiency of the student at that level and is taken as the panelist's estimate of the cut-score

Achievement Level Setting

- The procedure is repeated for each cut-score
- After all panelists have completed the task, results are shared and discussed in small groups
- Panelists then repeat the bookmarking exercise in light of the discussion
- Results are averaged in the group, and data are then provided as to what percentage of students would be classified into each proficiency level using the cut-scores obtained

Achievement Level Setting

- A third round of bookmarking may then take place, and cut-scores are averaged across all panelists to obtain the final cuts
- These cut-scores are then submitted to educational policy-makers for approval

Scaling and Score Reporting

- The proficiency scores produced by the psychometric model are on a scale with an arbitrarily defined zero point and no intuitive meaning
- A reporting scale is defined by rescaling the proficiency scores to some convenient or desirable metric
- Different states choose different reporting scales
- SBAC reports scores on a scale from 2000-3000
- The psychometric model provides a means for placing scores on tests in different grades on a common scale so that growth across grades can be measured

Scaling and Score Reporting

- Score reporting is the component of statewide assessments of most interest to educational administrators, teachers, parents, and the public
- Individual and aggregate score reports must be provided
- Score reports should provide clear information about how scores should be interpreted

Scaling and Score Reporting

- Guides are prepared for teachers, principals and administrators on the appropriate interpretations and uses of results for students
- These guides need to provide a description of the purpose and content of the assessments, show how to interpret the results, and specify appropriate uses and limitations of the data
- Reports and guides should use simple language that is understandable to parents, teachers, and principals

Validity Evidence

- Validity refers to the degree to which interpretations of test scores are supported by evidence (AERA/APA/NCME Test Standards)
- Primary sources of validity evidence:
  - Evidence based on test content
  - Evidence based on response processes
  - Evidence based on internal structure
  - Evidence based on relations to other variables
  - Evidence based on consequences of testing

Validity Evidence

- Evidence based on test content
  - Alignment studies are performed to show how the content of the assessment matches the academic content standards of the State
- Evidence based on response processes
  - If a claim is made that items measure certain cognitive processes, then evidence is required that those cognitive processes are required (often based on think-aloud protocols or examination of student reasoning)

Validity Evidence

- Evidence based on internal structure
  - If the test is assumed to measure a single construct, then statistical evidence should be provided that supports this assumption
  - Dimensionality analyses are routinely performed to investigate the presence of secondary dimensions that introduce irrelevant variance into test scores

Validity Evidence

- Evidence based on relations with other variables
  - Test scores should correlate more highly with alternate measures of the same construct than with measures of different constructs
- Evidence based on consequences
  - Example: evaluation of the effect of tests on instruction

Validity Evidence

- Other sources of validity evidence:
  - Careful test construction
  - Adequate measurement precision
  - Appropriate test administration
  - Appropriate scoring
  - Appropriate scaling and equating

Validity Evidence

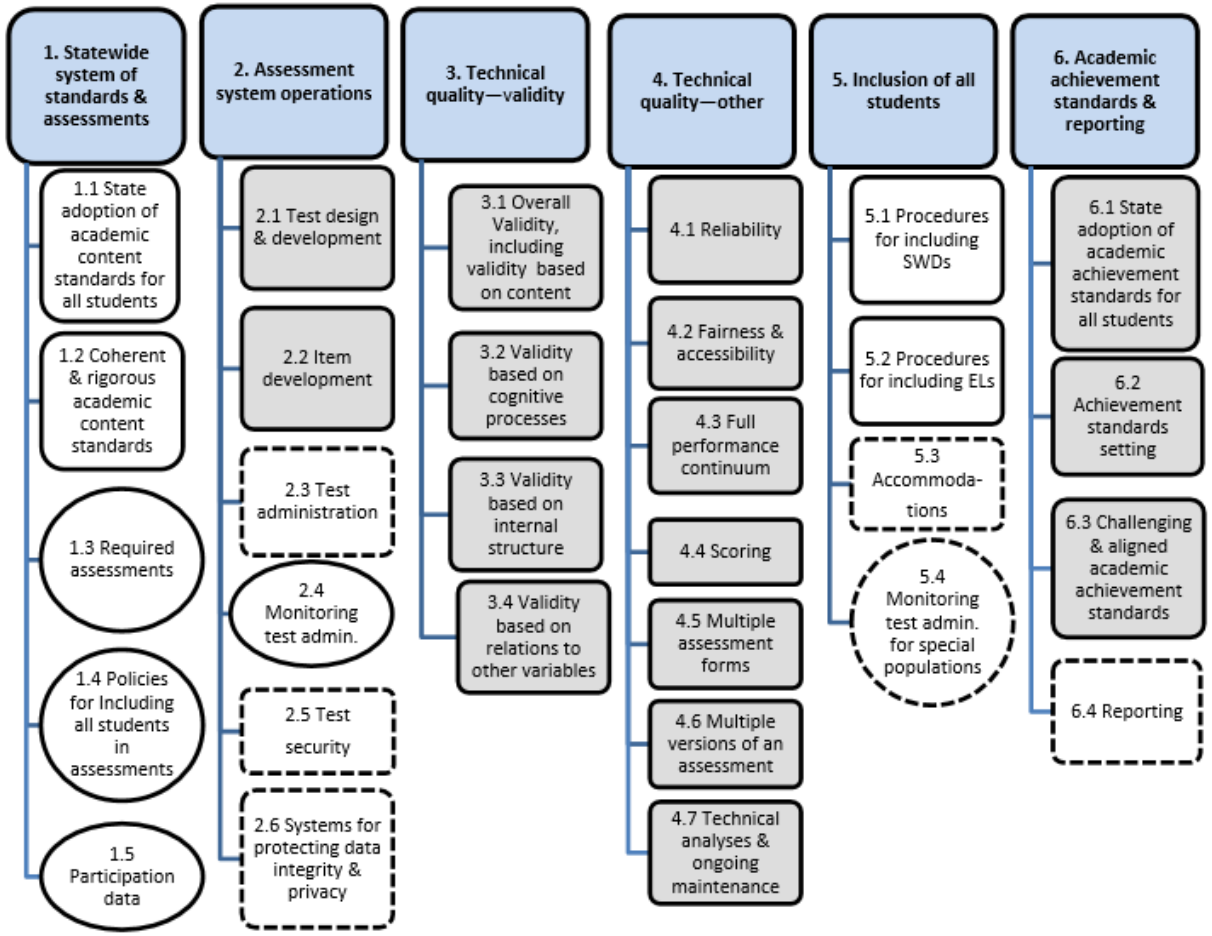
- Other sources of validity evidence:
  - Appropriate standard setting
  - Attention to fairness
  - Adequate test security

Critical Elements of a State Assessment System

- The steps outlined here are considered critical elements of a high-quality statewide assessment by the U.S. Department of Education, which recently released new guidance for state assessment systems to meet the requirements of ESSA
- State assessment systems will be reviewed according to a set of critical elements (U. S. Department of Education Peer Review of State Assessment Systems: Non-Regulatory Guidance for States, September 2015)

Critical Elements of a State Assessment System

Map of the Critical Elements for the State Assessment System Peer Review



KEY

- Critical elements in ovals will be checked for completeness by Department staff; if necessary, they may also be reviewed by assessment peer reviewers (e.g., Critical Element 1.3). All other critical elements will be reviewed by assessment peer reviewers.
- Critical elements in shaded boxes likely will be addressed by coordinated evidence for all States administering the same assessments (e.g., Critical Element 2.1).
- Critical elements in clear boxes with solid outlines likely will be addressed with State-specific evidence, even if a State administers the same assessments administered by other States (e.g., Critical Element 5.1).
- Critical elements in ovals or clear boxes with dashed outlines likely will be addressed by both State-specific evidence and coordinated evidence for States administering the same assessments (e.g., Critical Element 2.3, 5.4).