DOCUMENT RESUME

ED 270 473 TM 860 352

AUTHOR Legg, Sue M.; Algina, James
TITLE Practical Questions about Item Response Models in Large-Scale Assessment Programs.
PUB DATE Apr 86
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; Educational Assessment; Elementary Secondary Education; Equated Scores; Estimation (Mathematics); Goodness of Fit; *Latent Trait Theory; *Mathematical Models; *Scaling; Statistical Analysis; Test Format; *Testing Problems; *Testing Programs; Test Items; Test Reliability
IDENTIFIERS Test Content

ABSTRACT

This paper focuses on the questions which arise as test practitioners monitor score scales derived from latent trait theory. Large scale assessment programs are dynamic and constantly challenge the assumptions and limits of latent trait models. Even though testing programs evolve, test scores must remain reliable indicators of progress. Fundamental questions relate to the extent that score shifts may be due to changes in achievement or to the way in which achievement is measured. Over time a number of measurement concerns have been raised as item calibrations and score scales are monitored. These concerns are related to the effect on score scales due to item selection procedures and changes in the content of the tests. The following questions are discussed: (1) Can equating procedures accommodate changes in curriculum and test content? (2) What are the effects of variations in item format, population, and test administration? (3) What are the effects of different item difficulty distributions on score scales? (4) Which estimation procedure or latent trait model best fits the data? and (5) How can the meaning of test scores be enhanced? (PN)

PRACTICAL QUESTIONS ABOUT ITEM RESPONSE MODELS IN LARGE-SCALE ASSESSMENT PROGRAMS

by

Sue M. Legg and James Algina

University of Florida

Paper presented to the American Educational Research Association

San Francisco, California

April 1986

Because of item-free person measurement and person-free item calibration, latent trait models are potentially very useful in large scale assessment programs. However, these models require strong assumptions about dimensionality and the form of the item characteristic curves, and also require complicated estimation procedures. Large scale assessment programs, however, are dynamic and constantly challenge the assumptions and limits of these models. This paper focuses on the questions which arise as test practitioners monitor score scales derived from latent trait theory.

Of all the qualities that large scale assessment programs must have, perhaps the one most essential is resilience. The test development process for large scale state assessment programs does not produce a single examination for a singular purpose. Assessment programs may be designed to assess student achievement, evaluate and/or drive curriculum, compare individual schools or school districts, and even at times may be used to assess teacher effectiveness.

Particular goals change as political and educational environments change. Information about the improvement of students, schools and school districts on educational goals is used in the revision of goals and for allocating funding. Even though testing programs evolve, test scores must remain reliable indicators of progress. Yet, the scales on which these scores depend are subjected to constant pressure by changes in the curriculum and test specifications.

Testing programs in which passing standards have been set are subject to even more intense scrutiny. Small variations in the percent of students passing may affect large numbers of people. Declines in passing rates are viewed with alarm because of the implications for the

curriculum and the credibility of the schools. When these declines occur or even unanticipated increases occur, all possible explanations must be investigated. The fundamental questions relate to the extent that score shifts may be due to changes in achievement or to the way in which achievement is measured.

The interrelationship between the theory and practice of latent trait models provides new directions for research which can enhance the reliability and validity of latent trait measurement. For example, theory stipulates that tests must measure unidimensional traits. Practice dictates that tests include the measurement of skills, content, or format variations that may stretch this assumption. While latent trait measurement may be considered robust to the moderate violation of this assumption, little is known about the effect on item and ability estimates. Moreover, item selection practices may in themselves contribute to variations in item and ability parameter estimates.

Over time a number of measurement concerns have been raised as item calibrations and score scales are monitored. These concerns are related to the effect on score scales due to item selection procedures and changes in the content of the tests. We ask:

1. Can equating procedures accomodate changes in curriculum and test content?
2. What are the effects of variations in item format, population, and test administration?
3. What are the effects of different item difficulty distributions on score scales?
4. Which estimation procedure or latent trait model best fits our data?
5. How can we enhance the meaning of test scores?

Item Selection Procedures

Item selection may be accomplished by any one of several procedures.

For example, the following procedures have been used or considered in various Florida testing programs:

1. Items may be matched by skill and difficulty for each test form, or they may be drawn randomly from a pool of items representing each skill.

2. Specific numbers of items may be selected to measure each skill, or skills may be grouped in domains in which the number of items in a domain is fixed but the number representing each skill varies.

3. Skills may be fixed or drawn randomly, rotated or otherwise varied within content areas. In addition, the test blueprint may specify one item and skill selection procedure, and is later altered due to changes in curriculum requirements. Nothing appears to be known about how midstream changes in item selection procedures affect the stability of parameter estimates.

4. The distribution of item difficulties may be constant between test forms, or the average difficulty of the test may be fixed with variations in the item distributions.

Practices which permit between form variations in domains, skills, or item difficulty distributions give more flexibility in the selection of items from item banks, provide for more efficient item selection practices, and in particular permits a broader selection of items that represent each skill. However, little is known about the stability of parameters estimated from the results of forms constructed by the various item selection procedures.

Changes in the Content of the Test

A second test development occurence which relates to the dimensionality of examinations is a change in the content or format of an examination. Typical changes which occur are:

1. Item specifications are revised. The skills remain the same, but the way in which the skills are measured changes.

2. Tests measuring different cognitive levels of the same content are merged as in tests of basic skills and tests of the application of those skills.

There seems to be no universally accepted 'best' method for determining dimensionality. Correlation matrices derived from phi coefficients may indicate spurious factors related to item difficulty (Reckase, 1981). Matrices derived from tetrachoric coefficients have their own limitations, not only due to required sample sizes, but also to the assumptions concerning guessing.

Even the unit of analysis is open to question; should item scores or skill scores be correlated? In a recent report to the Institute for Student Assessment and Evaluation, Beard (1985) used skill scores to evaluate the dimensionality issues related to the possible merging of a test of basic skills and a test of the application of those skills. The rationale for this approach was given by Hulin, Drasgow, and Parsons (1983) who maintained that the use of skill scores avoided the correlational problems associated with dichotomously scored items.

Strategies to evaluate dimensionality by splitting tests into content areas as suggested by Bejar (1980) and cited by Hills and Beard (1985) have some practical drawbacks. State assessment programs develop multiple test forms each year which are drawn from large item banks. It is standard practice to evaluate dimensionality at the beginning of testing programs for initial test forms. Test developers often must operate on the assumption that a specific configuration of items which meet the specifications for the initial test is representative of every test form. As new forms are created in which skills are rotated, it is not feasible to conduct multiple factor analyses to examine dimensionality.

How can the equating process be structured to account for changes in curriculum and test content?

Equating with the Rasch model can be accomplished by several methods. The linking constant may be derived from a subset of items linked to other test forms in linear or triangular designs. A representative

set of items may be specified as a link, or all items in a form may be linked to a base or bank value. Items that do not fit the link are deleted. Either method is sensitive to the effect of items which fluctuate in difficulty. Estimates for certain skills or items within skills may fluctuate more than others. The addition or deletion of skills or items from the link can have a marked effect on the score scale.

As new items are calibrated and linked to the bank, their estimates are influenced by the particular set of items with which they are calibrated and by the limitations of the available sample for the field test. Base year values include the original items and items calibrated in subsequent forms. Base year values may be less stable than calibrations obtained after the testing program has been implemented for a period of time. Field test values for example, may have been obtained for test security reasons from students in teacher education rather from the intended population of teachers. When the curriculum has adjusted to the program and estimates become more stable, estimates may be averaged to create a new base value which should improve the linking process and improve the estimates for experimental items.

Item format, population, and test administration changes

In addition to the possibility that calibration and equating are affected by item selection procedures or changes in the skills measured by the test, one must be concerned about item format changes, population shifts, or test administration procedural changes; these all influence response patterns and are difficult to evaluate.

Some examples from recent studies conducted in Florida illustrate the issues which arise that relate to the structure and scaling of examinations. Data from one study indicated that increased testing time in mathematics may improve scores for hispanic students but not for black or caucasian students. If a change in testing time

were implemented, the interrelationship of the items may also change.

The reading subtest in one program consists of Cloze passages. The calibration of these passages may be conducted on an item level, or they may be calibrated as one item with multiple responses. There is some consensus that Cloze items in a passage are not independent and thus violate the assumptions of item level calibration methods.

Changes in methods of calibration have occurred in this reading subtest, and the format of the Cloze passages has also been revised. The stability of the item calibrations was examined when the length of the reading passages was increased to include twelve items per passage instead of ten with the foils decreased from three to two per item. An additional format change imbedded the responses with the passages rather than listing them on the margin. These revisions may alter the difficulty level of the items for people who speak English as a second language.

A third example of a practical issue with measurement implications relates to the calibration sample. Often a systematic random sample is drawn from the entire data set. Occasionally, district calendars or other problems prevent districts from complying with test administration deadlines. A test director may be confronted with drawing the calibration sample from an incomplete data set. If the test director proceeds with the scoring, then the standards for verifying the stability of the scale must be clearly defined.

Recently, a study was conducted in Florida to determine whether or not the scores from Dade county had any effect on the calibration sample. Dade county has a large immigrant population and significant numbers of

diverse minority groups. While the stability of estimates is not affected for different samples drawn from the same population, does the stability extend to the exclusion of large groups with potentially identifiably different response patterns?

Modifications to testing programs are constantly proposed; yet, it is seldom possible to have the time or research conditions necessary to evaluate these changes as part of the operational tasks of examination programs. Yet, these concerns which are often politically necessary should not be accomodated without knowledge of their measurement repercussions, but the evaluation criteria are unclear.

Item difficulty distributions

A plan for item selection using the Rasch item difficulties may require that items selected for different test forms have the same average difficulty level, and the difficulty estimates must be centered around the passing standard. A number of item selection models are possible under these constraints. Item difficulties may have normal, bimodal, or skewed distributions. When items are drawn from an item bank, extreme item difficulty values may be avoided because they make the task of creating equally difficult forms unwieldy. Thus, the item distributions may be truncated or elongated.

Another problem in item selection occurs when the passing standard set relatively high compared to the difficulty of the items as is often the case in basic skills examinations. The range of difficulty within the bank is fixed. As passing standards increase over time, the proportion of items with difficulty values below the passing standard increases. The addition of more difficult items to the bank may change the dimensionality of the test. What makes an item more difficult could be the skill, the format or even the foils. Given a broad range of

achievement in the population, a test centered at the highest passing standard may create face validity issues if not other validity issues.

While item difficulty and ability estimates are supposed to be invariant across samples of examinees drawn from the same population or over samples of calibrated items, little is known about the degree of inconsistency in parameter estimation which may be due to differences in item difficulty distributions or to lack of fit to the model. The extent to which these errors cause instability in the score scale and the passing rate must be evaluated. Current research in Florida on these questions (Beard, J. and Julian, E.; Legg, S. and Buhr, D.) has not provided a definitive answer.

A related issue is that as the curriculum changes, its assessment changes. Tests may become progressively easier, and/or more difficult versions of items may be introduced. The evolution of a testing program in which the ability distribution of the population changes may introduce scaling problems similar to those encountered for vertical equating.

Estmation Procedures and latent trait models

An issue also requiring investigation is the effect of the parameter estimation technique on the stability of the scale. Florida's testing programs use the joint maximum likelihood estimation procedure in BICAL. Other parameter estimation methods: marginal, conditional, or Bayesian may be more accurate depending upon the configuration of the data. Guidelines should be available for judging the stability of these different estimates under the varying data conditions described above.

Finally, while many studies have compared the efficiency of one, two and three parameter model estimation techniques, of particular interest is the comparative accuracy of these models for different distributions of data. Specific questions that have been addressed by Hills and Beard, 1985 in their report entitled, "An Investigation of the

Feasibility of Using the Three-Parameter Model for Florida's Statewide Assessment Tests" (SSAT) include:

...Will existing IRT computer programs work satisfactorily using the SSAT-II data?

...Is the assumption of unidimensionality valid for the SSAT-II data?

...Do the two- and three-parameter models fit the SSAT-II data better than the Rasch or one parameter model?

...Are the guessing (c) parameters estimable for the SSAT-II data using the LOGIST 4 computer program?

...How many examinees are needed to estimate the parameters?

...Can the parameter estimates be improved by oversampling the lower end of the ability distribution?

The SSAT-II is an easy examination with a highly negatively skewed score distribution. Hills and Beard found that the c parameter did not converge under the three parameter model. Fit for the two parameter model was better than for the one parameter model. Over sampling from the extremes of the distribution did not improve the estimations of the parameters.

How can the meaning of test scores be enhanced?

The multi-purpose nature of testing programs requires that as much information about student achievement be generated as possible. One recurring dilemma is the use of students' scores to both assess achievement and to provide data to identify student or curricular weaknesses. Skill level scores or content area scores are often requested both for individual students and for data aggregated by school, district or state levels. Frequently this information is reported as the percentage of students that correctly respond to the subset of items.

Percentage correct subscores are often not useful as indicators of achievement. They are subject to differences in the difficulty or

skills represented by the subscore items. One solution to this problem is to calibrate each skill or content section within a subtest separately, but there is research indicating that subscore ability estimates may overestimate the abilities estimated by the entire examination.

Rigid adherence to parallel item specifications and parameters negates the flexibility offered by test construction using item response theory. This request for more easily interpreted subscores should perhaps be rejected on the grounds that at best they represent a sample of items insufficient to use for diagnostic purposes and inadequate to indicate specific curricular weaknesses. Measurement experts may agree upon this rationale, but the political realities of statewide testing programs may demand othe alternatives.

The measurement community is beginning to respond to these concerns with development efforts in tailored testing. The Educational Testing Service has announced Project Jessica which will focus upon the construction of examinations which can pinpoint students' content, skill and problem solving deficiencies.

Summary

Established test development practices for norm referenced standardized examinations are somewhat different than those used for many state assessment examinations. Often multiple new test forms are developed each year from item banks. The banks themselves are constantly being revised and expanded. The format and content of examinations are also subject to frequent revision. The need to assess the stability of the score scales has real urgency. The pressure on statewide examination programs to respond quickly to policy changes requires that practitioners apply procedures to assess the limits of the models under varying conditions. Yet, these procedures are not well defined.

Bibliography

Beard, J. G. (1985) An Investigation of the Feasibility of Merging the SSAT-I and SSAT-II. Report to the Institute for Student Assessment and Evaluation, University of Florida.

Hills, J. R., and Beard, J. G. (1985) Investigating the Feasibility of Using the Three Parameter Model in Statewide Assessment Tests. Report to the Institute for Student Assessment and Evaluation, University of Florida.

Beard, J. G., and Julian, E. R. (1985) The Effect of Item Difficulty Distribution Shape on the Precision of Measurement at a Passing Ability. Paper presented at Florida Educational Research Association, Miami, Florida.

Bejar, I. J. (1980) A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurements, 17(4).

Hulin, C. L., Drasgow, F., and Parsons, C. K. (1983) Item Response Theory. Dow Jones - Irwin, Homewood, Illinois.

Legg, S. M., and Buhr, D. (1986) The Effect of Item Difficulty Distributiuon on Scaled Score Stability. Paper presented at the National Conference on Measurement in Education, San Francisco, California.

Reckase, M. D. (1979) Unifactor latent trait models applied to multifactor tests' results and implications. Journal of Educational Statistics, 4(3).