DOCUMENT RESUME

ED 294 901 TM 011 554

AUTHOR	Reckase, Mark D.
TITLE	Computerized Adaptive Testing: A Good Idea Waiting for the Right Technology.
PUB DATE	Apr 88
NOTE	19p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Latent Trait Theory; Test Items

ABSTRACT

The requirements for adaptive testing are reviewed, and the question of why implementation has taken so long is examined. The concept of a testing procedure that selects items to match the level of performance of an examinee during the administration of a test had to wait for the technology necessary to apply the idea. Current procedures were developed based on item response theory methodology. The reliability of shorter tests and scoring has been improved by this approach. Refinement of adaptive testing procedures is one aspect currently under development; a second is a focus on better ways to model person-by-item interaction and to produce test items to measure a person's skills. (SLD)

Computerized adaptive testing:
a good idea waiting for the right technology

Mark D. Reckase

The American Colle͜e Testing Program

Paper presented at the meeting of the American Educational Research Association, New Orleans, April 1988.

Adaptive testing has finally reached the point of operational implementation. Several large scale testing programs now use adaptive testing (Hsu & Sharmis, 1987; Knapp & Wise, 1987; McBride, Corpe & Wing, 1987; Ward, Kline & Flaugher, 1986), a commercial software system is now available for use in developing adaptive tests (Assessment Systems Corporation, 1984), and several possible implementations for adaptive testing are under investigation (Moreno, 1987; Olsen, Mayner, Slawson & Ho, 1986; Stevenson & Salehi, 1986). While it is gratifying to those of us who have worked in the area of adaptive testing for some time to see it finally reach the point of application, the question comes to mind: Why did it take so long? After all, standardized adaptive tests have been available since 1908 (Binet, 1908). This paper reviews the requirements for adaptive testing and suggests an answer to the question of why implementation has taken so long.

For many years, I have been observing the conduct of research in many different fields. These observations have led me to the conclusion that good ideas do not have much of an effect until the time is right for those ideas. The current theory of plate techtonics is a good example. For many years the idea that continents could move was thought to be silly. But eventually, after enough empirical evidence was accumulated, the theory of "sea floor spreading" became generally accepted. Similarly, the idea of adapting the difficulty of a test to each person tested in large scale testing programs had to wait until the time was right and the necessary technology was available before the reasonableness of the idea could be accepted and applied. It took approximately 80 years for this acceptance to occur, but now in 1988, the concept of adaptive testing is finally beginning to be adopted as a practical methodology within the set of procedures available to the measurement specialists.

Saying that the time was right for the concept of adaptive testing to be applied is not very informative, however. What changed in the way that test constructors thought of tests that allowed the transition to a new type of testing methodology to take place? The factors that led to the transition are critical to answering the question "Why now?" These will be enumerated in some detail after the components of an adaptive testing system are described.

Definition of Adaptive Testing

For the purposes of this paper, adaptive testing is defined as a testing procedure that selects items to match the level of performance of an examinee during the administration of the test. An operational adaptive test requires four components: a set of items from which the test is selected, a procedure for selecting the items, a method for computing a test score once the test is completed, and a means for determining when testing is done. The components of an adaptive test are described in more detail in Green, Bock, Humphreys, Linn & Reckase (1984).

Stanford-Binet as an Adaptive Test

In order to clarify the definition of adaptive testing, the 1960 edition of the Stanford-Binet (Terman & Merrill, 1960), a direct decadent of the Binet 1908 test, will be analyzed in some detail. Although the 1960 version will be used for convenience, the basic analysis applies to the earlier versions as well. A critical feature of the Stanford-Binet is that both the scores of examinees and the difficulties of test items are reported on the same score scale - the mental age scale (MA). The 1960 edition consisted of 122 tasks (items) arranged into 20 sets according to their level of difficulty. Items within a set were selected for the set because they were answered correctly about half the time by a particular age group. Thus, tasks in the six year set were answered correctly approximately half the time by six year olds. These 122 tasks composed the item pool of the adaptive test. During the test, items were selected to match the examinees ability by starting at a level in the 20 sets judged to be appropriate for the examinee and then administering easier tasks until a level was reached where all items were answered correctly (basal level) and then administering more difficult tasks until all items were answered incorrectly (ceiling level). It was assumed that all tasks below the basal level would have a 1.00 probability of a correct response and those above the ceiling level would have a 0.0 probability of a correct response.

The test was scored by adding a specified number of months for each correct item to the year designation for the basal level. This scoring procedure was in effect estimating the level on the mental age scale at which half of the items would be responded to correctly. The stopping rule for the procedure was to stop administering items when the basal and ceiling levels were determined.

More Recent Adaptive Tests

Although the Stanford-Binet type administration method had been in place for many years, it was not until the late 1950's to the early 1970's that attempts were made to adapt the test to the examinee on a larger scale than in one-on-one individual examinations. During that period of time, two-stage testing (Angoff & Huddleston, 1958), pyramidal testing (Krathwohl & Huyser, 1956), and the flexilevel test (Lord, 1971) were investigated. These procedures placed test items in a particular structural arrangement based on p-values and developed fixed paths through the items to match the test to the examinee. Weiss (1974) gives a good summary of these procedures.

The currently popular procedures for adaptive testing were developed during the late 1960's and early 1970's based on item response theory (IRT) methodology. These procedures used item pools that wer' precalibrated using item response theory models rather than group statistics, such as p-values. Items were selected for administration based on mathematical functions of the item-parameter estimates, such as item information or the minimum posterior variance of a Bayesian procedure. Scoring was also model based using maximum likelihood or Bayesian estimaticn procedures. Finally, the test was stopped when a decision was made, a level of precision was reached, or when a fixed number of items had been administered. Hulin, Drasgow & Parsons (1983) provide a summary of the procedures for IRT-based adaptive testing. For the most part, the previously developed structure-based procedures have lost favor to these new methods. The two-stage, pyramidal, and flexilevel testing procedures have never been used in operational testing programs. The next section of this paper will discuss each component of an adaptive testing system and indicate why the current procedures have supplanted the earlier methods.

Requirements for a Practical Adaptive Test

The Stanford-Binet was never used for large scale testing because of the requirement for one-on-one administration, the complexity of the administration of the items, and the time required for administration and scoring. Are there similar problems that led to the rejection of procedures like the pyramidal and flexilevel tests, or were the IRT based procedures simply better? Each component of an adaptive test will be considered to determine whether links to the reasons for the success of IRT-based procedures can be determined.

Item Pool

At the most basic level, the items in an item pool are not affected by the procedure used to select and administer them. The same items could be used with any of the available adaptive testing procedures. What has changed, however, is the type of information collected about an item. Current adaptive testing procedures are based on item statistics determined using item response theory procedures. These procedures make it relatively easy to put item statistics obtained from the responses of several different groups of people to different sets of items on the same scale so that they can all be considered in the item selection process. The result is that item pools of any required size can be produced by linking together separate calibrations.

Previous methodology tended to use group-based item statistics, such as p-values, to build the structured sets of items. Since p-values for different tests or different samples are likely to be nonlinearly related and to have inconvenient scale properties, it was not easy to combine sets of items into a single pool.

The IRT formulation of information also led developers to think about how much information should be provided at different points along the score scale. Consideration of item pool size and characteristics followed (see Patience & Reckase, 1979). Since previous procedures, with the noticeable exception of the Stanford Binet, used group statistics to form the item pool, the effects of item pool characteristics were not readily considered. The set of items were considered as a unit, not as single pieces that could be used to construct a pool with particular characteristics.

Item Selection Procedure

Most current adaptive testing procedures select te ꞉ items to maximize the information provided at the most recent estimate of ability. Withcut the current computer technology and the development of item response theory methodology this item selection methodclogy could not be implemented. It can be argued that the previous adaptive testing algorithms were attempting to do the same thing, but with less success. Although the Stanford-Binet administration procedure administers the most informative items, it also administers items that are not very informative by testing until both ceiling and basal levels are reached. The items at the ceiling and basal levels are too difficult or too easy and are therefore not very informative. Flexilevel testing tended to give uninformative items as the process continued because the items at the extremes of the difficulty range were used. Because of the use of uninformative items, more items were used than were necessary to obtain an ability estimate at a specified level of precision. Ireland (1976) showed that an IRT-based test using the Stanford-Binet item pool could shorten the test appreciably without losing reliability. De Ayala & Koch (1986) found the flexilevel test to require about twice as many items as a IRT-based adaptive test of equivalent reliability. The efficiency gained by the current procedures results from considering whether each item will add to the information provided by the testing process.

Scoring Procedure

The scoring of non-IRT based adaptive tests has always been a problem. Classical test theory does not readily deal with cases where different examinees get different items and even different numbers of items. IRT-based procedures readily produce ability estimates on the same scale after each item has been administered. This feature is a result of including both item- and person-parameters in the same model. The Stanford-Binet comes closest to the current adaptive testing scoring procedures because both items and people were scaled on the same mental age score scale. The other procedures that were developed prior to the use of IRT never solved the scoring problem in a way that yielded good statistical characteristics.

Stopping Rules

Since traditional tests or early adaptive tests had no good means for reporting scores on the same scale when different people took different numbers of different items, stopping rules seldom were required. Fixed length tests were the norm. The Stanford-Binet had a variable test length, but at the expense of administering items to each examinee that were too easy and too hard. The result was a lengthy and frustrating testing session.

Adaptive procedures based on item response theory have substantial flexibility in specifying stopping rules. Procedures have been proposed for stopping at a specified level of information, at a specified posterior variance for Bayesian procedures (Owen, 1975), or when a decision is made with a specified level of certainty (Reckase, 1980). Of course, fixed length tests can also be used. The IRT-based procedures allow the number of items administered to be closely tied to the requirements of test use.

Factors that Distinguish Current from Earlier Adaptive Tests

The differences between current and previous adaptive testing procedures given above suggests that certain factors facilitated the development of the current adaptive testing procedures. These factors are summarized below.

The Item as the Measurement Unit

The critical feature of adaptive testing is that the assessment of the skill level of an individual is the result of numerous interactions of the person and individual test items. Each item provides some amount of information about the skill level of the person, but all items do not provide equal amounts of information. This conception of the measurement process is a major breakthrough because it leads to the idea that items have characteristics that are independent of the group to which they are administered. Prior to the work of Lord (1952), items were described using statistics based on group performance. In fact, texts prior to Lord's seminal work gave very little advise about how to select items for a test. However, Gulliksen (1950, 392-393), at least, was aware of the need for relatively constant item characteristics and suggested that better item descriptors be a goal for future research.

Classical test theory functioned mainly at the level of the test score and its characteristics. Little consideration was given to the effect of particular items. The focus on the test score did not encourage researchers to match items to each examinee. If a group was well measured, individual measurement was also expected to be good.

Item Information

Once items were considered as individual tools for use in assessing a persons ability, the next question was, for what range of ability does an item give useful information about the person? The Stanford-Binet answered the question by using items that had a probability of correct response for the examinee that was neither 1.00 nor 0.00. This resulted in items being used over a fairly broad range of ability. This range was typically three to four standard deviation units wide. Current theory (see Recka:e & McKinley, 1984) suggests that the effective range of an item is only .7 to 1.8 standard deviation units wide. Thus, the Stanford-Binet was administering many items that did not provide much information, resulting in a low level of measurement efficiency.

Since most adaptive testing procedures currently used select items to maximize test information, the concept of item infcrmation is clearly important to adaptive testing. However, prior to the availability of IRT, the concept of information was unknown. There was no classical measurement theory analog to the information function. Item quality statistics were all based on group performance. Gulliksen (1950), for example, was surprised to find that the biserial-correlation-item-discrimination inoex changed with the ability level of the group used to compute the statistics because he considered an item to measure equally well over the entire ability range. Only with the use of IRT concepts can the range of effectiveness of an item be understood.

Scoring

A critical feature of adaptive testing is that the score obtained by an examinee is independent of the particular set of items given. Although this concept was clearly a part of all adaptive testing procedures, non IRT based procedures had difficulty developing a reasonable scoring scheme. For example, with flexilevel testing (Lord, 1971), the score on the test is the number of correct responses, plus .5 if the last response was incorrect. This scoring scheme would not yield comparable scores for flexilevel tests with different items. The score on a flexilevel test was basically equivalent to a raw score on a traditional test.

In order to truly free the scoring process from the set of items used, the item characteristics must be included in the same model as the person ability parameter. The Stanford-Binet achieves this by using the mental age scale to describe both item difficulty and test score. A five-year item was one that approximately fifty percent of five year olds' could answer correctly. This placement of items on the score scale allowed the scoring of any set of items. Until this same scheme was developed for generic adaptive systems, scoring independent of the item set could not be achieved.

Summary

It should be clear at this point that the existence of useful adaptive testing systems is closely tied to the development of item response theory. The Stanford-Binet was, in effect, using item response theory in 1916, but in such a restricted way that it could not easily be generalized. Chronological age acted as a surrogate for the ability scale, allowing for the scaling of both items and people on the same scale. Until IRT became available, the concepts underlying the Stanford-Binet administration procedure could not be understood, and therefore they were not considered to constitute a generalizable model. In a sense the Stanford-Binet administration model is the equivalent to the theories of plate techtonics mentioned earlier that were not considered viable until the supporting observational data were in place. In the field of measurement, it took about 80 years for the necessary theory to develop that supported the Stanford Binet type administration

Future Directions

Now that the power of item response theory has been realized and is taking hold in operational adaptive testing programs, have all of the basic measurement problems been solved? Of course not! The work has just begun. Current research on adaptive testing seems to be taking two basic directions. The first direction is related to refining the current methodology. Concerns over whether items function tne same on computer or in paper-and-pencil form (Divgi, 1986) or considerations of how to calibrate items as part of an adaptive test (Samejima, 1988) fall into these categories. These are critical areas of research, but they are not likely to result in major advances in testing.

The second type of research effort focuses on better ways of modeling the person-by-item interaction and of producing test items to measure a person's skills. Item response theory models are being produced that are nonmonotonic (Thissen & Steinberg, 1984), polychotomous (Sympson, 1986), and/or multidimensional (Reckase, 1985). In the future, adaptive tests will be based on a more accurate representation of the person-by-item interaction. Procedures are also being developed to generate items by computer to match the required item characteristics (Bejar, 1986). While these procedures have very limited capabilities now, if they can be enhanced in the future, adaptive tests will have essentialiy unlimited item pools that do not require calibration. The adaptive test of tomorrow may be equivalent to the idealized, infinite length tests of today.

Measurement has reached a new golden age. The number of interesting problems and promising approaches to solve them are almost limitless. But a key feature, which is an outgrowth of IRT methodology, is consideration of the test item. This is the equivalent for field of measurement of the discovery of the concept of the atom for chemistry and physics.

References

Angoff, W. H. & Huddleston, E. M. (1958). The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test. (Statistical Report SR-58-21). Princeton, NJ: Educational Testing Service

Assessment Systems Corporation. (1984). User's manual for the Micro CAT Testing System (ONR-85-1). St. Paul, MN: Author.

Bejar, I. I. (1986). Final report: adaptive assessment of spatial abilities. Princeton, NJ: Educational Testing Service.

Binet, A. (1908). Le developpement de l'intelligence chez les enfants. L'Année Psychologique, 14, 1-94.

De Ayala, R. J. & Koch, W. R. (1986, April). A computerized implementation of a flexilevel test and its comparison with a Bayesian computerized adaptive test. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Divgi, D. R. (1986). Determining the sensitivity of CAT-ASVAB scores to changes in item response curves with the medium of administration (CRM86-189). Alexandria, VA: Center for Naval Analysis.

Green, B. F., Bock, R. D., Humphreys, L. B., Linn, R. L. & Reckase, M. D. (1984). Tecnnical guideline for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Hsu, T. & Sharmis, M. (1987, April). Assessing psychometric qualities of an adaptive placement testing system. Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). Item response theory: application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons.

Ireland, C. M. (1977). An application of the Rasch one-parameter logistic model to individual intelligence testing in a tailored testing environment. Dissertation Abstracts International, 37(9-A), 5766.

Knapp D. J. & Wise, L. L. (1987, August). Refining the Computerized Adaptive Screening Test (CAST). Paper presented at the meeting of the American Psychological Association, New York.

Krathwohl, D. R. & Huyser, R. J. (1956). The sequential item test (SIT). American Psychologist, 2, 419.

Lord, F. M. (1952). A theory of test scores. Psychometric Monographs, 7, 1-84.

Lord, F. M. (1971). The self-scoring flexilevel test. Journal of Educational Measurement, 8, 147-151.

McBride, J. R., Corpe, V. A. & Wing, H. (1987, August). Equating the computerized adaptive editioı f the Differential Aptitude Tests. Paper presented at the meeting of the American Psychological Association, New York.

Moreno, K. E. (1987, August). Military applicant testing: replacing paper-and-pencil with computerized adaptive tests. Paper presented at the meeting of the American Psychological Association, New York.

Olsen, J. B., Maynes, D. D., Slawsn, D. & Ho, K. (1986, April). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the meeting of the Amerian Educational Research Association, San Francisco.

Owen, R. J. (1975). A Bayes'an sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Patience, W. M. & Reckase, M. D. (1979, April). Ope-ational characteristics of a Rasch model tailored testing procedure when program parameters and item pool attributes are varied. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Reckase, M. D. (1980). Some decision procedures for use with tailored testing. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference, 79-100.

Reckase, M. D. (1984, April). Item difficulty reconsidered: an IRT perspective. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9(4), 401-412.

Samejima, F. (1988, April). A robust method of on-line item calibration. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Stevenson, J. W. & Salehi, S. (1986, April). Project Adapt: an investigation of testing procedures for functional testing. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Sympson, J. B. (1986, August). Extracting information from wrong answers in computerized adaptive testing. Paper presented at the meeting of the American Psychological Association, Washington, D.C.

Terman, L. M. & Merrill, M. A. (1960). Stanford-Binet Intelligence Scale: manual for the third revision, form L-M. Boston: Houghton Mifflin.

Thissen, D. & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49, 501-520.

Ward, W. C., Klire, R. G. & Flaugher, J (1986). College Board Computerized Placement Tests: validation of an adaptive test of basic skills. (Research Report RR-86-29). Princeton, NJ: Educational Testing Service.

Weiss, D. J. (1974). Strategies of adaptive ability measurement. (Research Report 74-5). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.