Description of components in tailored testing

WAYNE M. PATIENCE
University of Missouri, Columbia, Missouri 65201

The major purpose of this paper is to describe the components of testing systems that address the primary goal of producing tests tailored to each individual's needs and abilities. The principal components of a tailored testing procedure are: item calibration, item selection, and ability estimation. Each of these components is defined and discussed. Alternative methods for each component are presented, along with an indication of their relative complexities. Finally, the particular methods used for each component in the tailored testing procedure at the University of Missouri-Columbia are described.

The major purpose of this paper is to describe the components of testing systems that address the primary goal of producing tests tailored to each individual's needs and abilities. Other goals of such testing systems include efficiency with regard to the use of student and instructor time, self-pacing, and flexibility in scheduling the session. These systems have been variously labeled adaptive testing, sequential testing, and tailored testing. Tailored testing (Lord, 1970) will be used in referring to these systems in this paper. Each makes use of the capabilities of current computer facilities and relatively recent developments in the area of measurement theory.

Many models used in tailored testing are based on latent trait theory, or latent structure analysis, as it was initially labeled (Lazarsfeld, 1959). Examples of these models include the normal ogive (Lord, 1952), the three-parameter logistic (Birnbaum, 1968), and the Rasch simple logistic (Rasch, 1960). Latent trait theory suggests that an individual's performance on some psychological task may be assessed by estimating his relative level on traits relevant to the given task. Tests are constructed to measure traits thought to comprise the ability being evaluated. Tailored tests, in addition to estimating the pertinent traits, also attempt to improve characteristics of the testing session.

Tailored testing systems are designed to select and administer to each individual some "best" set of items selected from a large pool. The primary components of a tailored testing procedure are: item calibration, item selection, and ability estimation.

ITEM CALIBRATION

The first step in setting up a tailored testing procedure is the calibration of test items for the item pool. Item calibration is the process of determining parameters that describe the characteristics of test items. Typically, item parameters indicate the difficulty, discrimination, and probability of a correct response when the student guesses. This information is used as a basis for selection of items for inclusion within the item pool, as well as selection of items for administration during the testing session of an examinee.

The type of item parameters estimated depends upon the particular method used for item calibration.

In addition to the traditional method of item analysis presented in most basic measurement and evaluation texts (see Sax, 1974), there are at least three other commonly used procedures for item calibration. One is the simple logistic model (Rasch, 1960). Another is based on the three-parameter logistic model (Birnbaum, 1968). The third is the normal ogive (Lord, 1952). Each of the item calibration models are adaptations of latent trait models. The normal ogive and three-parameter logistic provide difficulty, discrimination, and probability of guessing parameters. The simple logistic model, a special case of the three-parameter logistic, yields an easiness parameter and an indication of the item's goodness of fit. Each of the calibration procedures accommodates dichotomous items, that is, those items that may be scored either right or wrong.

Difficulty parameters of items calibrated using the Rasch simple logistic model are denoted by the log of easiness values. Increasingly positive parameters indicate less difficult items. Easiness values about zero describe items of midrange difficulty; the more negative easiness values denote items of increasing difficulty. The model states that when a person encounters any item, the outcome is solely a function of the person's ability and the easiness of the item. One of the Rasch model assumptions is that calibrations are independent of the sample of persons used to estimate item parameters, and ability estimates are independent of the selection of items used. The mathematical properties of object-free instrument calibrations and instrument-free object measurement (Rasch, 1966) provide the basis for item analysis from samples other than those to be tested with the tailored testing procedure. These properties also allow for comparisons of ability estimates measured on similar but not identical instruments, as well as adaptations of measurement tests to suit an individual's ability level. Each of the item calibration models discussed in this paper adheres to sample-free instrument calibration and instrument-free person measurement to some extent (see Wright, 1968). The Rasch model

assumes equal item discrimination and no guessing. The "goodness of fit" item statistic indicates to what extent an item meets the Rasch model assumptions. Those items that fail to meet the assumptions are not included in item pools for tailored testing.

The three-parameter logistic and normal ogive models yield the previously mentioned difficulty, discrimination, and guessing parameters. These parameters are derivative of item characteristic curves (Lord & Novick, 1968) that graphically represent the function of examinee ability and the probability of a correct response on the item at the levels of ability. The curve resembles the normal ogive (S shape) for both models. The difficulty of an item equals the ability level at which the item has a .5 probability of a correct response, assuming the item has zero guessing. The item discrimination is indicated by the slope of the curve at that ability level. The discrimination provides an indication of item quality with regard to the amount of information it provides about an examinee's ability. An item is good, or appropriate, to the extent that the probability of a correct response increases as the level of ability increases. The guessing parameter equals the probability of a correct response at the lowest ability levels.

The three-parameter logistic and normal ogive models differ in the functions employed to obtain the item calibrations and the complexity of the mathematical models, but the item parameters derived are quite similar. The three-parameter logistic is the simpler of the two and may be more readily applied. Rasch easiness parameters, when correlated with three-parameter logistic and normal ogive difficulty parameters, yield moderately high negative values (the scales of easiness and difficulty are reversed). Those who advocate use of the three-parameter logistic and normal ogive models suggest that the additional information provided by item discrimination and guessing parameters improves both selection of items for item pools and selection of items for administration to individual examinees. However, the additional information complicates the tailored testing procedure dramatically, as noted under "Item Selection." Evaluation studies of the relative merit of the calibration procedures have yet to be concluded.

ITEM SELECTION

Depending upon the procedures used for tailored testing (i.e., structure based or model based), methods of item selection for administration to an examinee may vary. Structure-based models have predetermined routing procedures through the item pool. Arrangement of items in the pool in combination with the routing method totally define the item selection process. Model-based procedures, on the other hand, are founded upon mathematical models such as the Rasch, three-parameter logistic, or normal ogive. For model-based procedures,

item selection is designed to maximize the information function. Arrangement of items in the pool is designed to provide efficient computer search for the item which would provide the most information.

The information function serves as the basis for determining which items from the pool are appropriate for administration to an individual examinee. The function provides a measure of each item's contribution to ability estimation at the different levels of ability. The information function is maximized for the Rasch model by selection of items that have a probability of a correct response of .5 for a given ability estimate. Initially, an item is administered from the pool with .5 probability of a correct response at an average level of ability. If the response is correct, the procedure selects a more difficult item. If the response is incorrect, an easier item is selected. Until at least one item has been missed and one answered correctly, the procedure has a fixed step size. Step size specifies the amount of increase or decrease in easiness parameters for item selection from one item administration to the next. There is considerable difference of opinion concerning the best step size to use. When at least one item has been answered correctly and one incorrectly, ability is estimated and an item is selected with probability .5 of a correct response at this ability level. If the exact item requested is not in the pool, the item with easiness nearest to the desired item is administered, unless the difference between the item requested and the nearest item is not within plus or minus the acceptance range. Acceptance range is an arbitrary value specifying the amount of deviation in the easiness an administered item may have from the requested item easiness without having adverse effects on the ability estimate. If no item is available within the acceptance range, the session is terminated. Note that item pool characteristics play a substantial role in the utility of a tailored testing procedure. The pool is sufficient to the extent that it can provide appropriate items over a broad range of ability. When substantial gaps between item easiness or difficulty parameters are present in the pool, there is propensity for ability estimate bias incurred from administration of inappropriate items or premature termination of the testing session.

Item selection procedures for the three-parameter and normal ogive models are considerably more involved. Item selection for administration to an examinee is again intended to maximize the information function, much as is the case for the Rasch model. However, selection of an item with .5 probability of a correct response based on item difficulty does not necessarily maximize the information provided with regard to the examinee's ability. Determination of that item which would provide the most information at a given point in the testing session must include consideration of discrimination and guessing, as well as difficulty parameters of every item in the pool each time an item is selected. Item selection becomes a

process of successively inserting all three parameter estimates for each item into the information function, to ascertain which item yields the highest value. The item determined to be the most appropriate for the ability estimate is selected and administered. The examinee's response initiates a new ability estimate if the response pattern includes at least one correct and one incorrect response, and the process is repeated. As is evident, the item selection for three-parameter and normal ogive models is considerably more complicated than for the Rasch simple logistic model. As a result, computer time is dramatically increased for each item selection for items with three parameters.

Major research questions with regard to the interaction of item selection procedures with the item pool and ability estimates are yet to be thoroughly resolved. What are the optimal step size and acceptance range values that minimize ability bias? Should the item pool be a normal or rectangular distribution of difficulty values? How few items can be administered by tailored tests without reducing the reliability and validity of ability estimates below that of comparable traditional tests? (For a discussion of these issues, see Reckase, Note 1, Note 2.)

To this point, the tailored testing procedures for item selection described have been indicative of variable-branching maximum-likelihood methods. Another alternative, Bayesian models, is frequently applied to item selection from pools calibrated with three-parameter logistic and normal ogive methods. The distinctions between these alternatives will be pointed out under "Ability Estimation."

ABILITY ESTIMATION

Procedures for ability estimation for tailored testing may be categorized into two broad groups. One group contains procedures that serve the previously mentioned structure-based procedures; the other group primarily serves the model-based procedures. As noted earlier, structure-based tailored testing is less flexible because administration of items is largely predetermined. Therefore, ability estimation tends to be less complex. Conversely, model-based tailored testing is complicated by the adaptations of the test to the individual examinee during the testing session. In order for this "tailoring" of the test to be accomplished, ability estimates must be made following the examinee's response to each item.

Two primary mathematical models for ability estimation have emerged in recent years for utilization in model-based computer systems: maximum likelihood and Bayesian. The models are quite sophisticated, and are based on different probability theories. Bayesian tailored testing procedures were derived by Owen and have been described several times (Jensema, 1974; Urry, 1971; Owen, Note 3). The variable-branching maximum-likelihood procedure has also been discussed (Lord, 1970; Reckase, 1974; Weiss, 1974). Several

variations of Bayesian and maximum-likelihood procedures exist. In this paper, the methods implemented for a tailored testing procedure are interactive. The items and their parameter estimates, the item pool, the item selection procedure, the ability estimate procedure, and the examinee all interact during the tailored testing session. The primary intent of all these procedures is to measure an individual's ability as efficiently and as accurately as possible.

The empirical maximum-likelihood estimate procedure obtains ability estimates using an iterative search for the mode of the likelihood distribution. This occurs when at least one item has been answered correctly and one incorrectly. The responses obtained form a response pattern which consists of a string of 0s and 1s, corresponding to incorrect and correct answers. The iterative search for the mode begins by computing the likelihood of the response pattern for the first ability estimate. Through a process of doubling and halving the first ability estimate, and computing the likelihood each time, the mode of the likelihood distribution is bracketed.

When ability estimates above and below the mode are found, the distance between the two estimates is divided into an arbitrary number of intervals. The likelihood is determined at the boundaries of each of the intervals. The interval with the highest likelihoods provides ability estimates above and below the mode. This process of computing ability estimates for upper and lower limits surrounding the mode of the likelihood distribution is repeated until the difference between the limits is less than a specified epsilon. At this point, an ability estimate is provided to the desired accuracy. The rate of convergence is dependent on the size of the intervals used in the iterative phases, as well as on the epsilon value. The best item is then chosen and administered. With the additional response in the response pattern, the procedure is repeated until the testing session is terminated. The session is concluded if the item pool is depleted of appropriate items, ability has been estimated to the specified accuracy, or a set number of items has been administered.

The Bayesian procedure for tailored testing is more complex than the maximum-likelihood approach. For this reason, it will not be presented in as much detail, but the major distinctions will be noted. When either of the procedures is used in combination with items having three parameters, ability estimation weights the items, depending on the amount of information provided by the items administered. Selection of items by either procedure is designed to administer the items from the pool which yield the most information about the individual examinee's ability. Bayesian procedures differ from maximum-likelihood approaches to ability estimation in the distributions assumed to be most representative of the examinee's ability at given points in the testing session. The Bayesian procedure assumes a normal distribution of ability at the onset of the

session. The parameters of the distribution are modified as the response pattern records correct and incorrect answers to the items administered. The cycle of administering items and estimating ability continues until the variance of the distribution is within a predetermined limit. Due to the differences in the probability theories of maximum-likelihood and Bayesian procedures, one could hypothesize that the number of items administered at different levels of ability are apt to be dissimilar. The maximum-likelihood procedure will probably administer more items in the midrange, while the Bayesian procedure will probably give more items for ability estimates at the extremes, all other factors held constant. This result may be proposed to be due to the fundamental contrast in the statistical camps with regard to probability theories.

EXAMPLE OF A TAILORED TESTING PROCEDURE

In the Educational Psychology Department at the University of Missouri-Columbia, a tailored testing system has been developed based on empirical maximum-likelihood estimation of ability from items calibrated with the simple logistic (Rasch) model. The following description of the system is an overview of the methods employed from each of the previously discussed components in an operational interactive approach to testing.

Item calibration data for the tailored testing system includes parameters from the simple logistic as well as three-parameter logistic models. Computer programs are used to obtain calibration data on items for the item pools of the tailored testing procedure. The programs are run on each of the exams in the undergraduate measurement course. Test items are initially tried on paper-and-pencil tests. On the average, 600 students take the senior-level measurement course in the 9-month academic year, and this provides good samples for item calibration. Item pools are constructed of items deemed to be of sufficient quality and content, based on the item calibration data. Presently, there are five item pools available for access by the tailored testing system. Items are arranged in the pools from the most difficult to the easiest, to allow efficient search during the item selection phase.

When an examinee is tested, the first item administered has a probability of .5 of a correct response for a person of average ability. If a correct response is obtained, the next item selected is more difficult. If an incorrect response is given, the next item is easier. The procedure has a fixed step size of .693 until a correct and incorrect response are present in the response pattern. At that point, an ability estimate is made using the maximum-likelihood procedure described earlier. Item selection switches from a fixed step size to item selection on the basis of .5 probability of a correct response for the ability estimated. At

present, the tailored testing procedure uses only the difficulty parameter for ability estimation and item selection.

If the item requested by the procedure is not in the pool, the item nearest to that parameter is selected, unless it is beyond ± 3 , the acceptance range. If no item is within this range, the session is stopped.

The procedure continues the interactive cycle of selecting and administering items, recording the response pattern, and making ability estimates, until one of the stop rules terminates the session. The stop rules include termination if the item pool has been depleted of appropriate items, ability has been estimated with sufficient accuracy, or 20 items have been administered.

Studies have indicated the tailored testing system to be a viable procedure. The reliability and validity of tests administered on the terminal are comparable to traditional tests containing about twice as many items (Reckase, Note 2). Advantages accrued with the tailored testing system include a reduction in time required to administer decreased numbers of items necessary for reliable ability estimation and more flexible scheduling of the testing session.

REFERENCE NOTES

1. Reckase, M. D. *The effects of item pool characteristics on the operation of a tailored testing procedure*. Paper presented at the spring meeting of the Psychometric Society, Murray Hill, New Jersey, 1976.
2. Reckase, M. D. *The reliability and validity of achievement tests administered using a variable branching tailored testing model*. Unpublished manuscript, 1976.
3. Owen, R. J. *A Bayesian approach to tailored testing*. Princeton, N.J.: Educational Testing Service, Research Bulletin, RB-69-92, 1969.

REFERENCES

- BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley, 1968.
- JENSEMA, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement*, 1974, 34, 757-766.
- LAZARSFELD, P. F. Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3). New York: McGraw-Hill, 1959.
- LORD, F. M. A theory of test scores. *Psychometric Monograph*, No. 7, 1952.
- LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley, 1968.
- LORD, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computerized assisted instruction, testing, and guidance*. New York: Harper & Row, 1970.
- RASCH, G. *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research, Copenhagen, 1960.
- RASCH, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in mathematical social science*. Chicago: Science Research Associates, 1966. Pp. 89-107.

- RECKASE, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods & Instrumentation*, 1974, 6, 208-212.
- SAX, G. *Principles of educational measurement and evaluation*. Belmont, Calif: Wadsworth, 1974.
- URRY, V. W. *Individualizing testing by Bayesian estimation*. Seattle: Bureau of Testing, University of Washington, 1971.
- WRIGHT, B. D. Sample-free test calibration and person measurement. *Proceedings of the 1967 invitational conference on testing problems*. Princeton: Educational Testing Service, 1968. Pp. 85-101.
- WEISS, D. J. *Strategies of adaptive ability measurement*. (Research Report 74-5). Minneapolis: Psychometric Methods Program, University of Minnesota, December 1974. (NTIS No. AD A004270).