DOCUMENT RESUME

ED 193 266 TM 800 581

AUTHOR Eisenberg, Eric M.; Bock, Cassandra L.
TITLE Applying Latent Trait Theory to a Course Examination System: Administration, Maintenance, and Training.
INSTITUTION Michigan State Univ., East Lansing. Dept. of Communication.
PUB DATE [80]
NOTE 24p.; For related document see ED 189 105.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Course Content; Equated Scores; Goodness of Fit; Group Testing; Heterogeneous Grouping; Higher Education; *Introductory Courses; *Item Banks; *Latent Trait Theory; *Test Construction; Testing Programs

ABSTRACT

Guidelines are described for setting up an item bank under latent trait theory which may be applied to the achievement testing system of multi-section, large-enrollment, college survey courses. The enrollment for the course is typically heterogeneous: students may be majors or non-majors, any one section may contain honors college students and disadvantaged learners, and students may be of differing class levels. The advantage is that test scores can be standardized without legislating a common examination for all sections. Unique findings of this study were: (1) the computer file containing serial positions of items on examinations proved extremely useful: (2) obvious violations of unidimensionality and asking questions based on trivial, unevenly taught information led to items which did not fit the model well: (3) less stability in the equating procedure may result when large differences occur in an examination system: (4) more than one training session for instructors is recommended: and (5) latent trait models can work in a large, multi-section course examination which surveys a variety of ostensively different topics. (RL)

APPLYING LATENT TRAIT THEORY TO A COURSE EXAMINATION SYSTEM: ADMINISTRATION, MAINTENANCE, AND TRAINING

Eric M. Eisenberg

and

Cassandra L. Book

Department of Communication
Michigan State University
East Lansing, MI

ABSTRACT

This paper is an attempt to bridge the gap between a somewhat complicated methodology and practice. Latent trait models are currently being applied to the achievement testing system of a large multi-section survey course at Michigan State University. The major objective of this paper is to provide educators of similar courses with practical guidelines for implementing a testing system which incorporates latent trait theory.

The more technical aspects of this implementation are treated in detail by Douglass (1980). Specific objectives of this paper include:

1. A description of the course examination system as it currently operates.
2. A rationale for the application of latent trait theory to an achievement test system.
3. A description of the steps necessary for organizing an item bank using latent trait theory.
4. Suggestions for the informative classification of items in the bank.
5. Considerations for maintaining the bank, with emphasis on criteria for the inclusion or exclusion of items.
6. A discussion of the training of instructors to use the bank.
7. Implications of latent trait theory for large multi-section courses.

DESCRIPTION OF THE COURSE EXAMINATION SYSTEM

Like many other courses with large enrollments, Communication 100 is a multi-section survey course covering a wide range of communication contexts, principles, and skills. The enrollment for the course is typically heterogeneous; students may be communication majors or non-majors and any one section of the course may contain honors college students and disadvantaged learners. Students of differing class levels (50% freshmen, 25% sophomores, 25% juniors and seniors) also take Communication 100.

Communication 100 is taught every quarter and has an enrollment of approximately 3000 students yearly. The course is divided into 15-20 sections each quarter with 50-70 students in each section. The course is overseen by a course director who attempts to ensure a standard quality of teaching and commonality of course content across all sections through training and periodic evaluation of section instructors. Each section is taught by a graduate student in communication who has primary teaching and evaluation responsibilities for that section. In this way instructors may determine the types of activities they prefer to use in teaching each of the topics, in addition to the order in which they determine best to cover them.

The topics covered in Communication 100 are related in that they each address an aspect of human communication, but differ in the setting for the communication or the specificity of focus. The ten topics covered each quarter range from public speaking to organizational communication, to the effects of the mass media.

Both a midterm and a final examination of the standard four-option multiple choice variety are administered each quarter. Each instructor prepares his or her own 40-45 item midterm examination. A special item pool is available which contains items used only on midterm exams. The final examination is created by the course director using 100 items from an item pool (derived from previous final examinations) and new items written by current instructors. An equal balance of items from each of the ten topics is included on the exam. The final exam in common and is administered to all sections simultaneously. Makeup examinations for mid-term and final examinations are drawn from their respective item pools.

Like other courses that test large numbers of examinees on a regular basis, Communication 100 test items are organized in a well-developed item pool. Each time is used a final examination it is typed on a 5"x7" index card and filed according to the ten topics outlined above.

In addition to the text of an item, the item pool contains:

1. Classical item statistics for each administration of the item on a final examination (difficulty is calculated as proportion of examinees getting the item wrong and the upper-lower 27% discrimination is used).

2. The date of each administration of the item.

3. Information about the effectiveness of each of the distractors (how many examinees in the upper 27%, middle 46%, and lower 27% selected each distrator).

4. An indication of when items were rewritten and the changes that were made, for referencing the appropriate set of item statistics.

INADEQUACIES OF CLASSICAL TEST THEORY IN TEST CONSTRUCTION

Although it is desirable to permit instructors to construct and to administer unique midterm examinations a comparability problem resulted. There is no straightforward way to compare examinees who had been examined with different subsets of items using classical test theory. Comparability is also a problem in the case of makeup examinations. There is no simple way to compare a student's score on a makeup examination with scores on the original exam when the two exams contain different items.

Further, classical item statistics recorded for each administration of an item often differed somewhat (as would be expected) across administrations. Since these statistics are used as the primary criteria for constructing equivalent tests, it is desirable to investigate item statistics which would be more stable across samples of examinees.

Although educators are most accustomed to interpreting classical item statistics, this approach has its limitations, as alluded to above. It has been observed by many researchers (Wood, 1976; Bejar, Weiss, & Kingsbury, 1977; Hambleton & Cook, 1977; Douglass, 1979) that the major

limitation of the classical model is lack of comparability of ability estimates and item statistics across different samples of items or examinees. These limitations are particularly relevant in this achievement testing situation; students may vary greatly in their ability (e.g., developmental students, upper and lower classmen, honors students), and instructors may construct examinations with quite different item difficulties and discriminations. An alternative psychometric basis for testing has been proposed which speaks to the limitations of classical test theory.

APPLICABILITY OF LATENT TRAIT THEORY TO THE ACHIEVEMENT TESTING SYSTEM

Latent trait theory (item response theory, item characteristic curve theory) has been offered as an alternative to classical test theory because it addresses the limitations described above (Wright, 1968; Rasch, 1960; McBride & Weiss, 1974; Lord, 1976; Hambleton & Cook, 1977). As Hambleton and Cook (1977) point out, the assumptions of latent trait theory are stronger than those for classical test theory, but strong assumptions imply strong results. When these assumptions are met reasonably well, one can expect (a) ability estimates which are independent of the sample of items or examinees chosen, (b) item statistics which are invariant across sub-groups of examinees, and (c) an estimate of the precision of ability estimation at each ability level (Hambleton & Cook, 1977).

The key assumption underlying latent trait theory has to do with unidimensionality of the latent space. Under latent trait theory, each item on any given test is measuring the same latent ability (in this instance, communication knowledge) as all other items. One should keep in mind when considering this assumption that practically it can never be met, and one is, in fact, attempting to assess how seriously this assumption can be violated while still obtaining stable estimates.

It is easier to see how the unidimensionality assumption could be defended for aptitude tests than for achievement tests. Items on aptitude tests often follow a central theme (test homogeneous content) and individual items which do not work well statistically may be discarded with little worry. The achievement test situation is different since test constructors typically attempt to construct an examination which is fair (covers most of what is taught) and balanced across topics in the course. Topics in a survey course can appear to be conceptually very different, as they do in Communication 100. In the achievement test context, it was our intent to determine the degree to which latent trait models are robust to violations of the unidimensionality assumption, and heterogeneous items can be included on an examination calibrated using latent trait theory. (For a complete discussion of how latent trait models were applied to this testing system, see Douglass (1980)).

The rationale for implementing latent trait theory in the achievement test context has been presented. The next section details the steps one takes in organizing a latent trait item bank.

Organizing the Latent Trait Item Bank

Some writers in the area of item banking make the distinction between item banks and item pools (Wood, 1976). An item pool is any collection of test items which serves as a resource for test construction. In contrast, Wood (1976) defines an item bank as "An all purpose measurement system capable of meeting any testing requirements, group or individual, and rooted firmly in latent trait or item characteristic curve theory." The test items for Communication 100 before this research began could best be classified as an item pool. One of the major purposes of this research was

to convert this item pool into a fully calibrated item bank, following Wood's definition.

In order to reap the benefits of latent trait theory outlined in the first section of this paper, it is necessary to make item statistics from all items in the pool vary along a common scale, and this is accomplished through item calibration. A convenient test is chosen to be the calibrating test (in this case, we chose the Spring 1979 final examination) and item statistics are calculated using one of the latent trait programs. In this study the one-parameter Rasch model was used via the BICAL program (Wright 1979). All of the items on the calibrating test are by construction on the same scale. To get the remainder of the items in the pool on to the common scale, items from the calibrating exam are included on new final examinations along with items which have not as yet been calibrated. It is recommended for precise calibration that approximately one-half of the items should be from the calibrating test. In addition, the calibrating test may be constructed items appearing on past test may also be calibrated.

Item parameters are estimated for those items not on scale. The transformation is found which places item statistics on the Spring 1979 scale. Douglass (1980) describes the mathematics of this process.

The above discussion has concentrated on the scaling of item statistics. Equating can be easily accomplished once scaling is done (Douglass, 1980; Lor, 1977; Wright, 1977). New exams may be equated as long as there

are additional items to place on scale. Each time a 100 item final examination is administered, forty of fifty new items may be added to the item bank. This is a gradual process; once all items have been calibrated, any subset of items drawn from the bank should yield the advantages discussed at the outset of this paper; ability estimates and estimates of item statistics independent of sub-groups of examinees and samples of items. To goal however, is not to develop a "closed" set of calibrated items, but rather to continually expand the back each term.

Once the equating process begins, there remains the logistical problems of recalling which examinations each item appeared on, which items have been calibrated, and which items appeared on the calibrating test. The next section describes the way in which items were classified in the item bank.

CLASSIFYING ITEMS

All of the information available in the classical item pool was retained in the item back. In particular, the following information was stored in the back:

Unique item number

Along with the text of each item, it was important to include a number which would uniquely identify it. This was to facilitate easy referencing of specific items which appeared on computer output, and to avoid confusion among tiems which appeared on different test but had the same number. The unique number had five digits, the first two of which identified the topic area, and the next three a specific position within that topic area. In Figure one, 25 indicates a question about nonverbal codes, and 032 is unique identifier.

Insert Figure 1 about here

Classical item analysis statistics

Classical item discrimination and difficulty were included to facilitate the transition between testing models. Traditional criteria for test construction were used for creating exams during calibration. Classical item statistics were useful in determining which types of items under classical statistics were best or worst under latent trait models. Information about how many people responded to each distractor was also included to aid in item revision.

Dates of previous adm stration and position

Each time an item was included on a final examination, the date of the examination and the serial position of the item on the examination were recorded. Referring again to the sample item, 784 indicates a Fall 1978 administration, and 791 a Winter 1979 administration. Twenty-four and 20 are the item's position on each of the two exams, respectively. If an item appeared on the Spring calibrating exam, this was marked with a red "S" for emphasis. (See Figure 1). Items which had appeared on exams which had been equated were marked with a green check.

A computer file was created which contained (a) the unique item numbers of all items in the back, and (b) the serial position of each item on each final examination of interest (in this case the four previously described). A portion of the computer file has been reproduced in Table 1. By looking at this file, one can tell immediately which items have appeared on the calibrating examination, which have not been calibrated at all, and where an item appearing on one of the examinations appears on the others.

For example, item 10057 in Table 1 appeared as item number 3 on the Fall 1978 exam, item 11 on the Winter 1979 exam, and item 12 on the Spring 1979 (calibrating) exam. This item was not used in the Fall 1979 exam. This provided a useful tool for calibration which bypassed the texts of items, and also a useful check on the number of previously calibrated items appearing on any given examination.

Insert Table 1 about here

Of the approximately 600 items available when this research began, 300 have been calibrated after five examinations of 100 items each. While our goal is to place all 600 items on a common scale, new items are constantly added to the bank through tests which contain at least fifty percent of their items already calibrated. Once an adequate number of items to draw different subsets of items have been calibrated, one can take advantage of the useful properties of latent trait theory outlined earlier.

Just as is the case with classical test theory, not all items will "fit" the model equally well. That is, certain items may not be adding very much to the measurement of the achievement one desires. The next section details the criteria used to decide which items to revise or omit from examinations.

MAINTAINING THE ITEM BANK

Once a good sized item bank has been established, one has the problem of deciding which items to include on an examination and which to exclude or revise. It seems apparent that one would use items which were (a) valid measures of the course content and (b) precise in their measurement of course content. The problem is not so clearcut. For example, the latent

trait model found to give the most stable estimates in this study works best for items of about average discrimination. It is common procedure to exclude items with low classical discriminations (e.g., below 20) since they do not add anything to the measurement. But what should be done with items having high (e.g., above 45) classical discriminations? Estimation programs may provide standard criteria for eliminating items which do not fit the testing model well. For example, Urry's (1976) procedure for the 3-parameter model does not report item statistics when latent trait discriminations are less than .80, difficulties less than -4.0 or greater than 4.0, or when the lower asymptote of the item characteristic curve is greater than .30.

In the case of the one-parameter Rasch model, high discriminators may be thrown out as well as low discriminators. While at first this appears to be sound measurement practice, one ends up removing many items which were considered best under the classical model. Since throwing out seemingly good items seemed wasteful, the researchers tried a different strategy. Total t-fit statistics were examined for each item to determine its "goodness of fit" with the latent trait model under consideration. For each final examination, the ten items having the worst t-fit were pulled from the bank. A sample of their classical statistics is reported in Table 2. Items removed by this process all had classical discriminations under 20 or over 45.

It seems a good policy to use evidence of poor fit as an indication that an item should be reviewed but not necessarily removed from the bank (or revised). About half of the items which were identified as having poor t-fits were seemingly good items; most of these were excellent under

classical criteria. Items which fit the Rasch model poorly but were good under classical criteria were used in subsequent test construction. The t-fit statistic is extremely dependent on the size of the sample of examinees. Also, t-fit is calculated relative to the items under consideration. Once low discriminators which fit poorly are removed, high discriminators which were poor fitters previously will fit the model better. In any case, t-fit statistics showed such instability across samples that it would be unwise to consider this standard "goodness of fit" criteria as the sole reason to exclude items from the bank or call for their revision.

Most of the items showing poor t-fit either (a) contained outdated information, (b) tested over trivial information, or (c) tested less important concepts which were not covered uniformly well across sections. These items will be revised and returned to the bank with a new unique number. Once an item has been revised it must be recalibrated. Even the slightest change in an item may make a difference in the way it is responded to.

Given the instability of goodness-of-fit tests, one is left with traditional criteria for including or excluding items. Further work needs to be done to develop stable goodness-of-fit tests for items. As under classical test theory, it appears to be a good strategy to exclude from the banks items which are (a) poor under classical standards, (b) have a clear grammatical or structural problem, or have no clear right answer, or (c) are invalid in terms of the course content, regardless of how well they fit the model.

The last criterion mentioned above is most likely to meet with opposition from some proponents of latent trait theory. In theory, once the

item bank has been fully calibrated, it doesn't matter which subset of items are used to construct tests; the resultant tests will be comparable. This is to say that one examination containing only questions from topic one will yield ability estimates for examinees comparable to an examination containing calibrated items concerning topic ten.

This argument may be more convincing in the case of aptitude tests, but seems faulty and misplaced when generalized to the achievement test context for the following reasons:

1. The stability of ability estimates from completely different subsets of items rests heavily on the unidimensionality assumption, which can never be entirely justified in practice. While stable estimates may result from tests which are somewhat balanced or overlapping in content, it seems unlikely that the model will be robust enough to yield stable estimates given this serious violation of the unidimensionality assumption.

2. More importantly aptitude and achievement tests are administered for different reasons. While an aptitude test may be concerned only with predicitve validity, an achievement test must cover a balance of topics taught to be fair. Even if the model was robust enough to permit comparable scores from dramatically different subsets of items, this would be inappropriate in an achievement test context.

It seems some difficulty may result when calibrating entire examinations. With all of the latent trait models, calibration to the Spring 1979 scale was based on the theoretical property that ability scales across tests differ only by a specified linear transformation (see Douglass; 1980). The Fall 1979 exam deviated more from linearity than either the Fall 1978 or the Winter 1979 tests[1]. While we were at first alarmed by this finding, it made more sense considering that in the Fall of 1979 there was a complete turnover of instructors and a substantially different training program. Although one

[1] correlation with the same items on the calibrating examination were .57, .80, and .87 for Fall 1979, Fall 1978 and Winter 1979 respectively.

instance is clearly not enough to make any strong claims, one can speculate that when radical differences occur in an examination system (types of students, instructors, or instructor training), less stability in the equating procedure may result. Since training may make a difference in the operation of these models, the next section details how instructors were trained under this system.

Training

A major concern of this study was that section instructors needed to be committed enough to this project to (a) write good test items which could be included in the item bank, (b) understand the basic assumptions of latent trait theory to facilitate their item writing, (c) understand the potential advantages of moving to a latent trait system to help motivate them, and (d) be willing to exert maximum effort in ensuring that tests were kept especially secure in the initial stages of calibration.

A two-hour workshop on item writing preceeded discussion of latent trait theory. Instructors were taught the basic skills for constructing items at different levels of cognitive learning as well as the common pitfalls of item writing.

Next, the potential advantages of latent trait theory were explained to the instructors. The new system was presented as a way to provide maximal instructor autonomy while allowing for comparable evaluation of students. It was emphasized that the item bank would still require well written, highly discriminating items. There was some discussion of the way in which widespread application of this system would be explained to students who might have questions or complaints about apparently unequal or different examinations given to different sections. Further information should be available about these concerns later this year.

As a part of their item writing training, instructors were informed somewhat about the unidimensionality assumption underlying latent trait theory. There has been a good deal of disagreement since the beginning of this study about whether or not common reading passage items (a series of questions relating to one stimulus) would violate the unidimensionality assumption, and this issue has not been resolved here. The concept of a valid item was introduced to instructors and examples were given of how an item might not be measuring what one would expect (but instead measuring vocabulary, or reading comprehension). More research needs to be done to determine which types of items are most likely not to fit the latent trait models due to serious violations of unidimensionality.

Finally, the necessity for security was emphasized. The researchers made it clear that while the loss of a test from a large item bank might not make a serious difference, the loss of a test at the early stages of calibration might lead to compromising one-fourth of the bank.

Although we conducted only one training session to orient and enlist the support of the instructors, it would appear that more frequent meetings would increase the likelihood of instructors writing effective new items. In addition, booster training sessions may also help to maintain morale and to encourage continued security in dealing with examinations. The instructors in this study were relatively inexperienced in item writing, and the results may have differed with more experienced instructors.

SUMMARY

This study has attempted to demonstrate the steps involved in setting up an item bank under latent trait theory. In sum, the steps are:

1. Decide whether you have a large enough course to meet the assumptions of latent trait models. Two-hundred students per term is probably the minimum using current estimation procedures.

2. Select an examination for calibrating which has as many items and examinees as is possible. Calculate item statistics for each item using one of the computer programs which executes latent trait estimation. We have found the Rasch (1966) model to be best and the BICAL program user-oriented (Wright, 1977). See Douglass (1980) for more technical information.

3. Nest items from the calibrating test on future tests and/or create a calibrating test which has many items in common with past tests. Do this until all items in the pool are on a common scale (and continue to add items).

4. Carefully record information about each item concerning (a) when it has appeared, and if it has been calibrated, (b) a unique item number for easy identification, (c) all information from classical item analysis (if available) to aid in the transition to latent trait models, and (d) the serial position of each item for each examination it has appeared on.

5. Once all items in the bank have been calibrated, one can expect the stable estimates which result from latent trait theory.

6. Avoid using items which have poor classical statistics, if this is possible. Do not use goodness-of-fit as an immutable criterion for excluding items; rather use it as a guide for determining which items to examine for potential problems. Include items which appear valid according to course content, unless their fit is very poor.

Unique findings of this study were:

1. The computer file containing serial positions of items on examinations proved extremely useful. Future plans include placing latent trait

statistics on the same file to aid in item selection and test construction.

2. Obvious violations of unidimensionality and asking questions based on trivial, unevenly taught information led to items which did not fit the model well.

3. When large differences occur in an examination system, less stability in the equating procedure may result.

4. More than one training session for instructors is recommended, particulary once test construction from the laten trait item bank is underway.

5. This study taken in combination with the work by Douglass (1980) shows that latent trait models can work in a large, multi-section course examination system which surveys a variety of ostensively different topics.

The guidelines described in this paper should prove useful to anyone involved with a large survey course with more than one section. Students and faculty alike have often felt discomfort in utilizing test results which are heavily dependent on the scores of other examinees or the particular subset of items which happened to appear on an examination in a particular quarter. The advantage of this system is that test scores can be standardized without legislating a common examination for all sections. Instructors retain the autonomy to arrange course content as they wish and to choose the test items which are most congruent with their teaching strategy or emphasis.

There are many other advantages of adopting a latent trait item bank. Makeup tests would provide results which are easily comparable with the original test. If groups with specific ability ranges are tested, examinations can be constructed to measure with most precision at the desired ability level. Under the latent trait models, estimates of

measurement error are available at all points along the ability scale, whereas classical test theory provides only an overall estimate. Finally, tests could be tailored for individual students and still result in fair comparisons of scores. Much worthwhile research in this area has already begun.

25032

784/ 791
24 20

Which of the following networks provides for the most democratic decision-making?

a. completely-connected
b. chain
c. wheel
d. circle

Fall 1978				Spring 1979				
DIFF 28 DISC 45				DIFF 28 DISC 41				
Options				Options				
0	1	2*	3	0	1	2*	3	
9	95	110	65	3	57	190	12	(red)S

Figure 1. Layout of a typical item in the bank.

Unique #	Fall '78	Winter '79	Spring '79	Fall '79
10057	3	11	12	
10058	6	10	6	91
10059				
10060	12	9	19	
20000				8
20001				
20002	4	22	50	

Table 1. A sample of the computer file referencing the position of each item on tests.

Unique #	Difficulty	Discrimination
30081	42	52
25009	55	52
10021	36	53
25007	45	49
20001	50	47
50019	43	17
80014	47	15
50013	35	7
40032	42	9
50015	62	-6

Table 2. A listing of classical item statistics for ten of the worst fitting items using total t-fit as the criterion.

REFERENCES

Bejar, I.I., Weiss, D.J., and Kingsbury, G.G. Calibration of an item pool for the adaptive measurement of achievement. (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977.

Douglass, J.B. A process for testing a mathematical model for the solution of a practical problem: applications to test equating. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 8-12, 1979.

Douglass, J.B. Applying latent trait theory to a classroom examination system: model comparison and selection. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts, April 7-13, 1980.

Hambleton, R.K., and Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

McBride, J.R., and Weiss, D.J. A word knowledge item pool for adaptive ability measurement. (Research Report 74-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1974.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institut, 1960.

Urry, V.W. A five year quest: is computerized adaptive testing feasible? In C.L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. Washington D.C.: U.S. Civil Service Commission, 1976, pp. 97-102.

Wood, R. Trait measurement and item banks. Advances in Psychological and Educational Measurement, New York: John Wiley & Sons, 1976.

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Inivitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968, 85-101.

Wright, B.D. and Mead, R., and Bell, S. BICAL: calibrating items with the Rasch model. (Research Memorandum No. 23B). Chicago: University of Chicago, Statistical Laboratory, Department of Education, 1979.