Symposium

The Development of an Item Bank in Mathematics Using the Rasch Model

David F. Robitaille
university of british columbia

Thomas O'Shea
simon fraser university

Item banks for use in the development of tests keyed to the provincial mathematics curriculum were developed at the Grade 3/4, 7/8, 10/11 levels. The Rasch simple logistic model was used to calibrate all of the items within each of the three banks on a common scale of difficulty. A computerized data base was developed to store the items, to facilitate modification of the item banks as required, and to produce camera-ready tests consisting of items meeting criteria determined by the user. To date the item banks have been used to produce eleven tests published by the Ministry of Education and as a source of items for the 1981 B.C. Mathematics Assessment.

Les banques d'items servant à l'élaboration de tests qui se rapportent au programme provincial d'études en mathématique ont été établies pour les 3e et 4e années, 7e et 8e années, 10e et 11e années. Le programme logistique simple de Rasch a été utilisé afin d'établir une échelle de difficulté commune à chacune de ces banques et de classer chacun des items en fonction de cette échelle. On a également établi un programme d'ordinateur pour entrer ces données en mémoire afin de pouvoir les modifier au besoin et produire des tests sur écran présentant des items capables de satisfaire les attentes de l'usager. Jusqu'à ce jour, ces données ont servi à élaborer 11 tests qui ont été publiés par le ministère de l'Education et qui ont servi de référence pour "the second B.C. Mathematics Assessment."

One outgrowth of the Learning Assessment Program in British Columbia (Mussio & Greer, 1980) has been an increased demand for valid and reliable achievement tests keyed to the prescribed curriculum for use by classroom teachers. Requests have come from individual teachers as well as from administrators at both the school and school district levels. As a first step in responding to such requests, the Mathematics Achievement Test (MAT) Project was set up to construct item banks and to develop tests in mathematics at three levels in the school system: Grade 3/4, Grade 7/8, Grade 10/11.[1]

The development of collections of test items for educational applications appears to be gaining in popularity both here and abroad. In British Columbia, the MAT Project has been expanded to include other grade levels, and similar items banks have been developed at several grade levels for measuring achievement in science. In Ontario, a great deal of work has gone into the development of the Ontario Assessment Instrument Pool. Elsewhere, the International Association for the Evaluation of Educational Achievement (IEA) has established an item banking project, and a number of member-countries of that organization have begun work on the development of a test-item bank in mathematics.

The development of item banks has been facilitated by recent progress in the area of data base management systems and by the use of latent trait models for calibration of items. Two of the major objectives of this article are to illustrate how these factors contributed to the completion of the MAT Project, and to illustrate how the system can be used to produce tests according to specifications determined by the user. Preceding this discussion will be a description of the procedures used to develop test items included in the item bank.

ITEM DEVELOPMENT

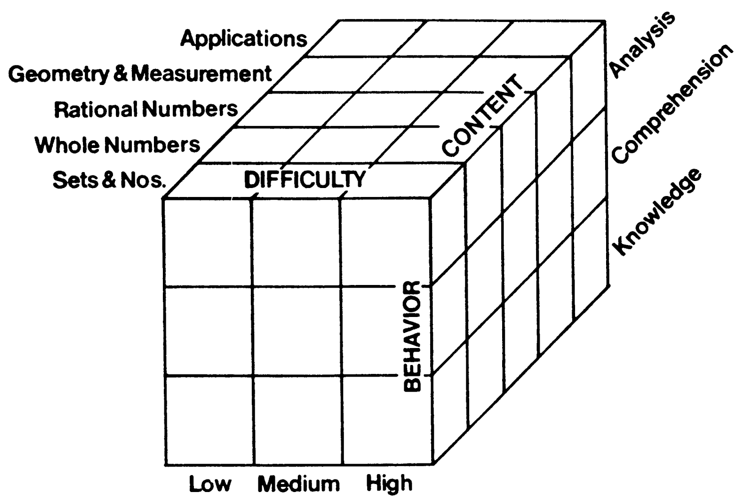
The three levels for which item banks were developed correspond to three major divisions in the mathematics curriculum in British Columbia: Grade 3 marks the end of primary education; Grade 7, of elementary school; and Grade 10, of compulsory courses in mathematics. Each item in the item banks may be used on tests designed to serve as summative measures of achievement or on tests designed for diagnostic or placement purposes. Thus the Grade 3/4 item bank may be used to construct achievement tests at the Grade 3 level or diagnostic tests for Grade 4 students.

In order to facilitate the process of item development, an Item Specification Model was developed for each level. As is shown in the Grade 7/8 model displayed in Figure 1, the three dimensions of the model were Content, Difficulty, and Behavior.

The first dimension of the model is mathematics content, and the official curriculum guide for mathematics (Curriculum Development Branch, 1978) was used to define those areas of content to be included. The curriculum guide contains lists of *learning outcomes* to be used by teachers in deciding topics to be taught. The separate outcomes are grouped into a number of different *strands* at each grade level and, as is shown in Figure 1, items were developed for five content strands for Grade 7/8. For Grade 3/4 three strands were used: Sets and Numbers, Operations with Whole Numbers, and Geometry and Measurement. For Grade 10/11 three strands were used as well: Algebra, Geometry, and Consumer Mathematics. Each strand dealt with a major area of content for the grade level concerned.

In the curriculum guide the learning outcomes are stated in fairly general terms (e.g., "the student is able to write sets of equivalent fractions"). In order to ensure that item writers agree about the precise meaning of each objective, *amplified objectives* (Popham, 1974) were

FIGURE 1
MATP Item Specification Model – Grade 7/8



produced. These amplified objectives specified, in some detail, the kinds of items admissible for each objective, and how those items should be constructed.

The second dimension of the model is item-difficulty. Since the item banks were intended for student use at all levels of ability, it was necessary that the items be of varying levels of difficulty. For each objective, item writers were asked to produce items of low, medium, and high difficulty in the ratio 2:2:1. These difficulty levels were defined to correspond to *p*-values of 80–100%, 30–79%, and 0–29% respectively. Judgments concerning item difficulty were made on an *à priori* basis by the item writers based upon their experience in mathematics teaching.

The third dimension of the model is level of cognitive behavior. Three levels were used: Knowledge, Comprehension, and Analysis, with 50% of the items for a given objective to be at the Knowledge level, and 25% at each of the other two. Definitions of these levels were adapted from Wilson (1971).

The actual task of writing test items was carried out by teams of writers for each grade who wrote the items in week-long sessions during which they were released from their regular duties. Each team included classroom teachers, supervisors, and faculty members. Two item writers were assigned to each strand and worked as a pair. Each item produced was checked three times by different people to verify the content categorization, as well as the assigned levels of difficulty and cognitive behavior. All of the items which survived this process of validation were prepared for field testing and calibration.

The original plans for the project called for the inclusion of both multiple-choice and open-ended test items. However, results of the pilot tests revealed serious problems with the interpretation of results on the open-ended items so they were dropped from the item banks. As a result, all of the items are multiple-choice items with four foils, each one a derivable response to the item. Responses such as "none of these" were not utilized.

THE RASCH MODEL

One of the requirements of the project was to investigate the feasibility of applying latent trait calibration procedures to develop item banks and the tests derived from them. The attractiveness of latent trait theory lay in two specific claims. First, it has been claimed that by calibrating items on a common difficulty scale, student scores on tests which contained no common items might be directly and meaningfully compared though the tests were not constructed to be equivalent in the classical sense. Second, it has also been claimed that the measurement of student achievement at a particular level can be made more precise by the judicious selection of items from a restricted range of difficulty. The MAT Project was designed to produce tests which would yield the least error of measurement at the Pass/Fail level since errors in characterizing student performance at this point were considered more serious than at other grading locations.

Two latent trait models were investigated for their potential application in this project: the three-parameter model, and the one-parameter model. In the three-parameter model, the probability of correct response by a person of given ability is assumed to be a non-linear function of the difficulty of the item, the item discrimination, and the probability of correct response through guessing. The model's theoretical complexities lead to practical difficulties, particularly in evaluating the suitability of procedures used to estimate the parameters.

The one-parameter, or Rasch, model is simpler since the performance of an individual of given ability is assumed to be a function only of the difficulty of the item. Based upon the demonstrated success of the Rasch model in various applications, particularly an item banking project in Portland, Oregon, it was selected for use in the MAT project.

In the period since the inauguration of the MAT Project, a considerable amount of controversy has arisen over the application of latent trait models. In Canada, discussion of the Rasch model was featured prominently at the 1981 meeting of the Canadian Educational Researchers Association in Halifax. In England, the criticisms of Goldstein (1979, 1980) have led to a reconsideration of the appropriateness of using the Rasch model as a means of monitoring changes in patterns of educational achievement over time. Researchers and others contemplating use of latent trait theory in applied situations should be acquainted with both sides of the argument before reaching any final decision concerning its use. For a general introduction to the area, Hambleton's (1979) presentation is recommended. A more detailed review of latent trait models and their applications, including specific suggestions for improvement, is contained in Traub and Wolfe (1981).

The one-parameter model used is a modification of one first proposed by Georg Rasch (1966). The probability function upon which the model is based is given by

$$P(si) = \frac{e^{a(s)-d(i)}}{1 + e^{a(s)-d(i)}}$$

where, $P(si)$ is the probability that person s correctly answers item i,

$a(s)$ is the ability of person s,
$d(i)$ is the difficulty of item i, and
e is the base of natural logarithms.

According to the Rasch model, person-ability and item-difficulty are related in such a way that if a highly capable person is faced with a low-difficulty item, the probability that that person will obtain the correct answer is very nearly 1.0. Similarly, in a situation where a person is faced with an item that is neither too difficult nor too easy for him, the outcome is uncertain and the probability of obtaining a correct answer is about 0.5.

The computer program BICAL (Wright & Mead, 1978) may be used to estimate the difficulty, $d(i)$, of each item and the ability, $a(s)$, of each person. All persons obtaining the same raw score, r, are assumed to be of equal ability. The common scale underlying both $a(s)$ and $d(i)$ is an equal-interval scale ranging, in theory, from negative infinity to positive infinity. The unit of measure for each parameter is the *logit* and, in practice, the usual range is approximately -4 to $+4$ logits. Negative values for $a(s)$ indicate low-ability persons, and negative values for $d(i)$ indicate easy items.

There are several advantageous properties of the model which have been investigated and generally supported through the analysis of data (Rentz & Rentz, 1978). First, within reasonable limits, the estimates of item difficulty are independent of the sample of persons used to calibrate the items. This is in direct contrast to the traditional model which uses the *p*-value, that is the percentage of correct response in the sample tested, as an index of an item's difficulty.

Second, an estimate of a person's ability may be obtained from any collection of previously calibrated items. Within reasonable limits the abilities of two persons may be assessed using completely different items, and the estimates of their abilities will be directly comparable. This characteristic is particularly desirable in the construction and utilization of item banks where the objective is to match items to persons. Using calibrated items, tests appropriate to individuals can be constructed, and ability estimates compared on the common scale.

Closely related to the foregoing is the matter of the standard error of the ability estimates. In the Rasch model the best estimate of a person's ability is obtained from a test consisting of items each with a success probability of 0.5. Hence, if the intention is to create a test which most accurately measures the abilities of low achievers, items having correspondingly low levels of difficulty should be selected from the item pool.

ITEM CALIBRATION

The fundamental requirement in the creation of an item bank is the determination of the difficulty of each component items on a common scale. By definition an item bank consists of a large number of items and, since each student in a calibration sample cannot be expected to respond to all possible items in the bank, a procedure making use of different students' responses to different sets of items must be employed. It is here that the power of the Rasch model is realized. By utilizing blocks of common items as links between different tests, the difficulty values of the items may be interrelated.

Different linking procedures were used at each of the three grade levels for which items were developed. The calibration procedure for the Grade 10/11 items is described in some detail here as a comprehensive example.

Three content areas or strands were defined for Grade 10/11: Algebra, Geometry, and Consumer Mathematics. Initially, the items in each strand were calibrated independently using the linking network shown in Figure 2.

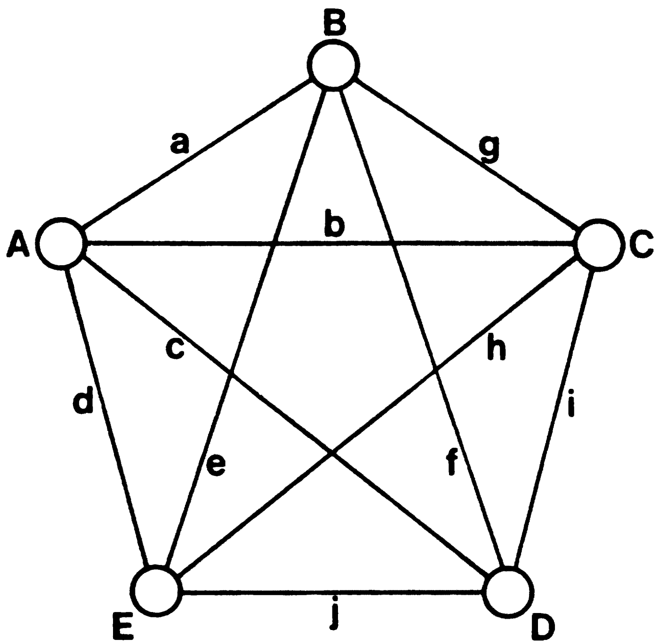
A, *B*, *C*, *D*, and *E* represent five test booklets each containing 40 multiple-choice items. Each booklet consisted of four blocks or links of ten items each. Each of the ten blocks of items ($a, b, c, \ldots, j$) linked two booklets. In all, 100 items were administered in ten links of ten items for each of the three strands. Sixteen test booklets were constructed: five for each of the three strands plus a sixteenth which contained items from all three strands.

For each strand, the responses of 1500 students were obtained from a provincial probability sample of Grade 10 classes by selecting an appropriate number of classes stratified by geographic location and size of school. At the same time the teachers of those students were asked to rate their students' achievement in mathematics by assigning them grades of *A*, *B*, *C*, *P* (Pass), or *F*. This information was used later to interpret the raw scores obtained on MAT tests in terms of letter grades.

The responses to the items in each booklet were first analyzed using traditional item analysis techniques (Nelson, 1974). Based on these results, some items were excluded from any further consideration, while others were set aside for possible future modification and calibration.

The remaining items were then subjected to Rasch calibration using BICAL (Wright & Mead, 1978). After the item analysis and Rasch

FIGURE 2
Linking network for items in a Grade 10/11 strand



calibration were completed, the numbers of items retained in the three Grade 10/11 strands were 92, 86, and 89 respectively.

Next the links were used to bring all calibrated item difficulties to a common scale. Figure 2 shows booklets *A* and *B* connected by the ten items in link *a*. Since *A* and *B* contain items of varying difficulty, and since the booklets were calibrated independently, the difficulty of each item in link *a* assumes two values: one resulting from its inclusion in booklet *A*, and one from its inclusion in booklet *B*. However, because of the equal-interval scale underlying the model, one set of difficulty values may be taken as the standard, say those from booklet *A*, and each value from booklet *B* may then be adjusted by the same amount to bring all of the items in both links onto the difficulty scale defined by booklet *A*. The size of the linking constant was further adjusted by considering other indirect links, that is, from *A* to *B* through *C*, through *D*, and through *E*. The result was three independently calibrated sets of items: Algebra, Geometry, and Consumer Mathematics.

The items which made up the sixteenth booklet were selected from all three strands, and they were used to standardize the difficulty scale across strands. Scaling constants were determined by comparing the mean difficulties of the items in each strand, thereby permitting all of the items from three strands to be calibrated on a common scale.

At all three grade levels, the original logit scale was transformed into a more convenient one, in which the centre of the scale was set at 1000 and each logit was set equal to 100 units. This new unit of measure was named the BRIT (*BR*itish Columbia un*IT*). Item difficulties on the BRIT scale range from approximately 700 to 1200.

THE DATA BASE MANAGEMENT SYSTEM

All of the more than 1200 calibrated items and their associated data elements were entered into a specially designed SPIRES data base. SPIRES (*S*tanford *P*ublic *I*nformation *RE*trieval *S*ystem) is a data base management system which utilizes the hierarchical data model (Berryman, 1980). At UBC, SPIRES runs under MTS (Michigan Terminal System) on the Amdahl 470 V/8.

Each piece of information about an item in a SPIRES item bank is called an *element*, and the collection of all of the elements for a given test item is a *record*. Figure 3 shows the form in which a typical record is stored in the data base.

The SPIRES record contains the item itself as well as a good deal of information concerning it. The grade level at which the item is to be used (GRADE), the type of item (TYPE), where the item came from (SOURCE), the correct answer (ANSWER), the topic (STRAND), and the cognitive level of behavior at which the item is directed (LEVEL) are all included, along with an indication of the objective from the official curriculum guide (DEPT-CODE) corresponding to the item.

Statistical information obtained at the item-calibration stage is also contained in each record. The *p*-value and point-biserial correlation coefficient are given for each answer choice, along with the Rasch difficulty score expressed in BRITs (RASCH). Finally, the record includes information about the extent to which the item has been used. For example, the element

STRANDS-TST = 1–0679–18

indicates that this item was used once on a test published by the Ministry of Education in June 1979, and that it is number 18 on that test. Depending upon the nature of the item, the record may also contain an element which indicates space required for the insertion of a diagram or figure.

SPIRES has a rich command language which gives the user a great deal of flexibility in updating the data base and in conducting searches of it. The data base can be searched for all or a portion of those items which fit the particular criteria desired for inclusion on a test. Any element can be used as one of the parameters of such a search, and elements may also be combined or have limits set on them for admissible values. As an example, the search request

FIGURE 3
Sample item as stored by SPIRES

```
REC01 = 6
DATE-ADDED = 01/23/79
ITEM-TYPE = Multiple Choice
SOURCE = Local
DATE-ACCESSED = 05/30/79
ITEM = " Last baseball season, John went up to bat 30 times.  In"
ITEM = "    those 30 times at bat, John got 18 hits.  What percent"
ITEM = "    of John's times at bat resulted in hits?"
ANSWER = b
GRADE = 7
  LEVEL = Analysis
  STRAND = AP - Applications
  REFERENCE-NUMBER = 07A006
  PILOT-TEST-REF = APa8
  DEPT-CODE = VII.2
  RASCH = 943
  P-VALUE-CORRECT = 59.90
  P-VALUE = 2.90
  P-VALUE = 59.90
  P-VALUE = 10.10
  P-VALUE = 19.60
  CORRELATION = -0.14
  CORRELATION = 0.42
  CORRELATION = -0.13
  CORRELATION = -0.23
  CALIBRATION-DATE = 78-10
STRANDS-TST = 1-0679-18
FOIL
CHOICE = " 167"
FOIL
CHOICE = "  60"
FOIL
CHOICE = "  30"
FOIL
CHOICE = "  18"
```

FOR GOAL WHERE GRADE = 3
AND STRAND = WN
AND LEVEL = KNOWLEDGE
AND RASCH > 800

produces 47 items from the Grade 3/4 bank, items which deal with operations with whole numbers (WN) at the level of computational skills and knowledge of facts, and with RASCH scores greater than 800 BRITs. If needed, this collection of items could itself be the object of a subsequent search designed, for example, to eliminate items corresponding to a certain curricular objective.

Output from such a search can be printed in any of six different forms using FORMAT definitions especially constructed for the MATP data base. For example, the FORMAT called TEST.OUT may be used to produce a camera-ready copy of a test containing the items selected in the search described above. TEST.OUT numbers the items, sets up and numbers the pages, and arranges each item in a standard multiple-choice

FIGURE 4
Sample item in TEST.OUT format

```
1.      Last baseball season, John went up to bat 30 times.  In
        those 30 times at bat, John got 18 hits.  What percent
        of John's times at bat resulted in hits?

                        a.  167 ..................[ ]

                        b.   60 ..................[ ]

                        c.   30 ..................[ ]

                        d.   18 ..................[ ]
```

format including boxes in which students can indicate their responses. Figure 4 shows how the item mentioned earlier would appear when TEST.OUT was used to control the format of the output.

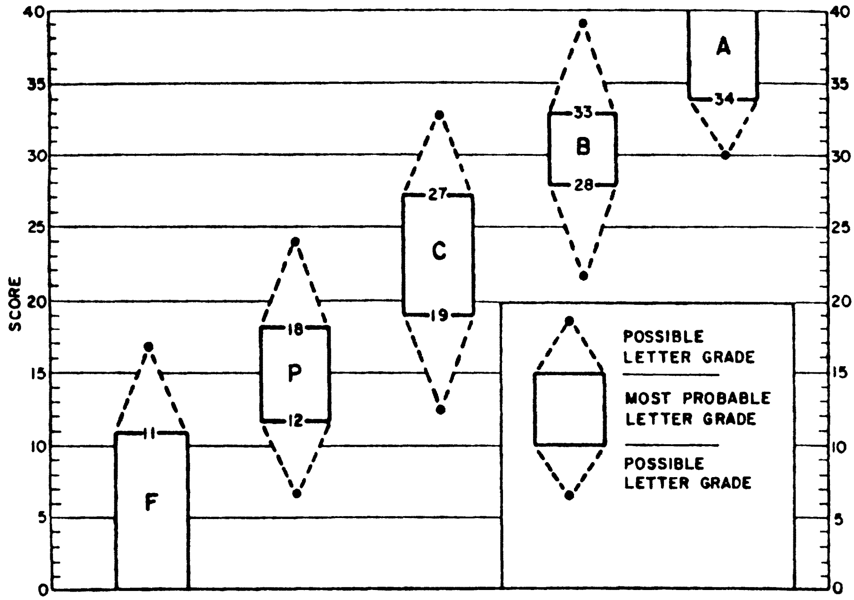
PRODUCTION OF STRAND TESTS

The item banks were used to produce eleven 40-item *strand tests,* one for each of the content-by-grade-level strands mentioned earlier. Each test was accompanied by a teacher's manual which included a graph such as the one shown in Figure 5, to enable teachers to convert raw scores on a particular test into letter-grade categories: *A, B, C, P,* or *F.*

Forty items of relatively low difficulty, providing the best coverage of the learning outcomes concerned, were selected from a pool of approximately 100 Rasch calibrated items to compile a strand test. This was done to increase the accuracy of what was considered to be the most crucial grading decisions at the Pass-Fail level.

Since the 40 items contained on each test had never been administered as a single group before, it was necessary to establish a procedure for determining raw score boundaries between letter grades. To this end, the raw scores of all students in the Grade 7/8 calibration sample were converted into BRIT ability scores and adjusted to a common scale. These were aggregated into a cumulative percentage frequency distribution as shown in Figure 6. By superimposing the percentage of students falling into letter grade classifications as indicated by the teacher responses, the critical values in BRITs were determined for the four letter grade boundaries. These ability scores were then used to produce a table of raw score/letter grade equivalents unique to each strand test.

FIGURE 5
Raw score to letter grade conversion chart. (Grade 7/8 – Sets and Numbers)



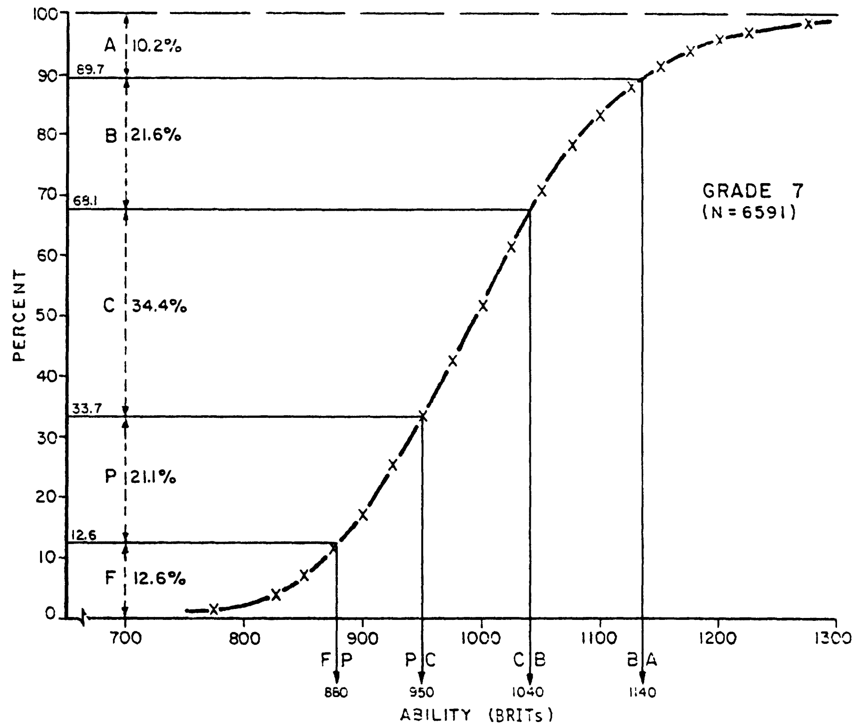
To assess the validity of the foregoing procedure, each of the Grade 7/8 strand tests was administered to a probability norming sample of 1000 students. On two of the five Grade 7/8 tests the raw score intervals for each letter grade suggested from the calibration sample and those obtained from the norming sample were identical. On the third, the A/B boundary was the same while the remaining boundaries differed by one unit. On both of the remaining tests, the boundaries differed by one or two units.

TABLE 1
Letter Grades & Raw Scores for 2 Strand Tests (Grade 7/8)

	Raw Score Intervals	
Letter Grade	*"Sets & Numbers"*	*"Operations with Whole Numbers"*
A	35–40	37–40
B	29–34	32–36
C	22–28	26–31
P	16–21	21–25
F	0–15	0–20

FIGURE 6

Use of the cumulative frequency distribution to establish grade cut-off scores



In no case was the difference greater than the standard error of measure associated with each cutoff score.

Table 1 gives some indication of the variation in raw scores corresponding to letter grades for different tests.

On balance, the intervals determined from the calibration sample appeared reasonable. A similar procedure, without benefit of the validation procedure employed at the Grade 7/8 level, was used to establish letter grade cutoff points for Grades 3/4 and 10/11.

The potential benefits to be derived from this type of test development procedure are considerable. It may be possible, for example, to eliminate the norming phase from the usual sequence of test development activities. With a large collection of calibrated items from which to draw, tests tailored to meet specific needs may be produced on very short notice. This is true particularly where the items are stored in a data base such as the one designed for this project.

CONCLUSION

The 11 strand tests published by the Ministry of Education are being widely used throughout the province. Schools ordered more than 200,000 copies of the tests during the first year in which they were made available. The item banks have also been used as an item source for the 1981 B.C. Mathematics Assessment. Finally, item banks for Grade 5 and Grade 9 mathematics as well as for science have been constructed using the Rasch model for item calibration.

NOTE

[1] Funding for this project was provided under the terms of a contract with the Learning Assessment Branch of the Ministry of Education, Province of British Columbia. Other members of the team which carried out the project were James M. Sherrill of U.B.C.; Heather Kelleher, a Learning Assistance specialist at John Robson School in New Westminster; and John Klassen, department head for mathematics at Sutherland Secondary School in North Vancouver.

REFERENCES

Berryman, J. Database management. *University of British Columbia: Computing Centre Newsletter*, 1980, *12*(10), 3–4.

Curriculum Development Branch. *Mathematics curriculum guide: Years one to twelve.* Victoria, B.C.: Ministry of Education, 1978.

Goldstein, H. Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 1979, *5*, 211–220.

Goldstein, H. Dimensionality bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 234–246.

Hambleton, R.K. Latent trait models and their applications. In R.E. Traub (Ed.), *New directions for testing and measurement: Methodological developments* (No. 4). San Francisco: Jossey, Bass, 1979.

Mussio, J.J. & Greer, R.N. The British Columbia assessment program: An overview. *Canadian Journal of Education*, 1980, *5*(4), 22–40.

Nelson, L.R. *Guide to LERTAP use and interpretation.* Dunedin, New Zealand: University of Otago, 1974.

Popham, W.J. Selecting objectives and generating test items for objectives-based tests. *Problems in criterion references measurement*, CSE Monograph Series in Evaluation, Number 3. Los Angeles: Center for the Study of Evaluation, 1974.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, *19*, 49–57.

Rentz, R.R. & Rentz, C.C. *Does the Rasch model really work? A discussion for practitioners.* Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, E.T.S. 1978.

Robitaille, D.F., Sherrill, J.M. & O'Shea, T. J. *Mathematics achievement test project: Technical manual.* Victoria, B.C.: Ministry of Education, 1980. (ERIC ED 187 531)

Traub, R.E. & Wolfe, R.G. Latent trait theory and the assessment of educational achievement. In D. Berliner (Ed.), *Review of Research in Education*, 1981, *9*, 377–435.

Wilson, J.W. Evaluation of learning in secondary school mathematics. In B.S. Bloom, J.T. Hastings & G.F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill Book Company, 1971.

Wright, B.D. & Mead, R.J. BICAL: Calibrating items and scales with the Rasch model. *Research Memorandum No. 23A*, Statistical Laboratory, Department of Education, University of Chicago, 1978.

David Robitaille is a Professor of Mathematics Education in the Department of Mathematics and Science Education, University of British Columbia, Vancouver, British Columbia V6T 1Z5.
Thomas O'Shea is an Assistant Professor in the Faculty of Education, Simon Fraser University, Burnaby, British Columbia V5A 1S6.