Banking Non-Dichotomously Scored Items

Geofferey N. Masters
University of Melbourne

John Evans
Development Dimensions International

A method for constructing a bank of items scored in two or more ordered response categories is described and illustrated. This method enables multistep problems, rating scale items, question "clusters," and other items using partial credit scoring to be calibrated and incorporated into an item bank, and it provides a mechanism for computer adaptive testing with items of this type. Procedures are described for calibrating an initial set of items, for testing the fit of items to the underlying measurement model, and for linking new items to an existing item bank. The method is illustrated using items from the Watson-Glaser Critical Thinking Appraisal.

The first requirement in the construction of an item bank is a collection of items. These items usually are organized by content and catalogued for easy reference. The second requirement is a psychometric method which can be used to construct a coherent measurement system from an item collection. This psychometric method is used to calibrate items, to identify items which are anomalous in their operation, and to provide measures which are freed of the particulars of the items attempted. Since the development of psychometric models capable of supporting such a measurement system, item banks have been established at a number of sites throughout the world, including Australia (Cornish & Wines, 1977; Hill, 1985; Tognolini, 1982), England (Choppin, 1968, 1976, 1978, 1981; Elliott, 1983), Scotland (Pollitt, 1979, 1985), the United States (Forster & Ascher, 1977; Koslin, Koslin, Zeno, & Wainer, 1977; Wongbundhit, 1985; Wright & Bell, 1984), and Austria (Kubinger, 1985).

A limitation of most calibrated item banks is that they contain only dichotomously scored items—usually dichotomously scored multiple-choice items. In some contexts this is not a serious limitation, but in others, the restriction of an item bank to only items scored correct or incorrect can place an unacceptable constraint on the forms of assessment the bank makes possible. This is particularly true of banks used in the assessment of educational achievement. For the assessment of skills such as problem solving and essay writing, dichotomously scored items usually are considered inadequate.

This paper describes and illustrates a method for banking test and questionnaire items scored in several ordered response categories. The measurement model applied here is a member of the Rasch family of measurement models, and thus permits person parameters to be conditioned out of the procedure for calibrating items (Rasch, 1960, 1977). This is especially desirable in the construction of an item bank because different items are invariably calibrated on different groups of individuals. Two uses of this method—to calibrate a core of items to begin an item bank, and then to link new items to an existing bank through common-item equating—are illustrated below.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 4, December 1986, pp. 355–367

0146-6216/86/040355-13$1.90

Polytomously-Scored Items

Items scored in several ordered response categories are widely used in educational and psychological measurement. They are particularly common in psychological questionnaires used to measure attitudes and personality; but partial credit scoring is also very common in tests and examinations scored by hand. If the advantages of item banking and computer adaptive testing are to be extended to items of this type, then a psychometric method for calibrating and banking items scored in more than two ordered response categories is required. This paper begins by outlining several different types of polytomously-scored items that might be calibrated and incorporated into an item bank. Although the data analyzed in this paper are based on one particular way of defining a set of ordered response categories, the method described here should be general enough for use with a wide variety of item types.

Multistep Problems

One example of an item which is usually not scored dichotomously is the multistep problem. Items of this type are common in educational assessment, particularly in subject areas such as mathematics and the physical sciences. These items are designed to assess students' abilities to identify an appropriate solution strategy and to pursue this strategy to a successful conclusion. It is usual in scoring items of this type to identify several intermediate stages in the solution of each problem and to award credit on the basis of the number of steps completed. In this way, several ordered levels of performance are defined for each item.

Although problem-solving items are common on educational tests and examinations, their inclusion in calibrated item banks is hampered by the limitations of most item banking procedures to dichotomously scored items. Bank developers wishing to include problem-solving tasks in their banks are confronted with the choice of either scoring multistep problems dichotomously so that they can be calibrated, or simply collecting and cataloguing these items without attempting to incorporate them into their measuring systems.

Rating Scales

Another common format for recording students' performances on an item is to rate responses on a scale (e.g., 1 to 10). This format is particularly popular for recording performances on tasks such as building a model, assembling a piece of apparatus, carrying out a procedure, or writing an essay. In assembling a collection of items of this type, each task (e.g., each essay topic) might be accompanied by a scoring scheme describing the criteria to be applied in rating performances on that task, perhaps with examples of attempts at each of the rating points. This entire set of materials (task, scoring criteria, examples) might then be treated as an "item" and calibrated for inclusion in a bank.

Items from attitude scales and personality inventories which provide respondents with a set of ordered response alternatives (e.g., Never, Sometimes, Often, Always) might also be calibrated and included in an item bank. Once calibrated, these items can be used in the construction of new questionnaire forms or as a basis for computer adaptive assessment (Koch & Dodd, 1985).

Question Clusters

Occasionally, test and questionnaire items come in clusters with all items in a cluster relating to the same piece of text. An example of a cluster of items, taken from the Watson-Glaser Critical Thinking Appraisal, is shown in Figure 1. The three items shown here (Items 65, 66, and 67) all refer to the sentence in italics immediately above them. Respondents are asked to evaluate the strength of each of these three arguments and to describe each argument as either "strong" or "weak." (Respondents are instructed not to let their personal attitudes toward an issue influence their evaluation of the argument. For an argument to be strong it must be both important and directly related to the topic.) Each response is then scored correct or incorrect.

One issue that arises in the use of clusters of this type is that of local independence. If Items 65, 66,

Figure 1
An Example of a Cluster Item

Should all young men in the United States go to college?

	Test 5 Argument Strong	Weak
1. Yes; college provides an opportunity for them to learn school songs and cheers. (This would be a silly reason for spending years in college.)	1 ☐	■
2. No; a large percent of young men do not have enough ability or interest to derive any benefit from college training. (If this is true, as the directions require us to assume, it is a weighty argument against all young men going to college.)	2 ■	☐
3. No; excessive studying permanently warps an individual's personality. (This argument, although of great general importance when accepted as true, is not directly related to the question, because attendance at college does not necessarily require excessive studying.)	3 ☐	■

When the word "should" is used as the first word in any of the following questions, its meaning is, "Would the proposed action promote the general welfare of the people in the United States?"

EXERCISES

Would a strong labor party promote the general welfare of the people of the United States?

65. No; a strong labor party would make it unattractive for private investors to risk their money in business ventures, thus causing sustained large-scale unemployment.

66. Yes; differences between Republicans and Democrats today are not as great as the difference between liberals and conservatives within those parties.

67. No; labor unions have called strikes in a number of important industries.

and 67 in Figure 1 are analyzed as three separate items (e.g., using the dichotomous Rasch model), then the assumption of local independence must be made. Each person's response to any one of these items must be assumed to be uninfluenced by his/her responses to the other two. In most dichotomously scored tests, this is a reasonable assumption. But within an item cluster, items have a shared dependence on a common stem and are thus less likely to be locally independent.

A second issue arises when items of this type are used in an adaptive test. The usual practice in adaptive testing is to select from an item bank the unused item that provides the most information about the person being measured. If the items in a cluster are treated as independent dichotomously

scored items and selected in the usual way, then a respondent may be required to read a long piece of introductory text (perhaps as long as one or two paragraphs of detailed information) to answer only one item in a cluster. If, later in the test, that person is directed to another item in the same cluster, then he or she may find it necessary to reread the text before attempting this second item.

In an adaptive test it is probably more efficient to treat a cluster as a single "item." If each cluster can be calibrated and stored in a bank, then the next most informative cluster can be selected for presentation. In this way, respondents are required to read the accompanying text only once and can attempt all questions in a cluster together. The score on each cluster is simply a count of the questions in that cluster answered correctly, and takes values between 0 and m (where m is the number of questions in the cluster).

Interactive Items

A fourth type of polytomously-scored item that might be calibrated and included in an item bank is a computer-administered item which provides feedback to respondents during a test. This feedback may simply inform respondents of their success or failure on each item and offer a second attempt if an item is failed. Failure on a second attempt might be followed by a third or fourth attempt. Under this "answer-until-correct" format (Whetton & Childs, 1982; Wilcox, 1982), credit on each item is awarded on the basis of the number of attempts required to provide the correct answer to that item.

In an interactive test, feedback might also be given in the form of one or more hints (Trismen, 1981, 1982, 1983). Under this format, persons failing on their first attempt receive a hint and an opportunity to try again. Failure after a hint may be followed by further assistance. Each person's score is then based on the number of hints required to arrive at the correct answer. Masters and Adams (1985) have investigated the use of latent trait methods to calibrate computer-administered items with hints.

The Partial Credit Model

The method developed here for banking nondichotomously scored items is based on a measurement model for ordered response categories. This model requires the a priori specification of a set of possible scores $(0,1, \ldots, m_i)$ that can be obtained on each item i. The model probability of person n with ability β_n scoring x $(x = 0,1, \ldots, m_i)$ on item i is denoted π_{nix} and is given by

$$\pi_{ni0} = \frac{1}{1 + \sum_{k=1}^{m_i} \exp\left(k\beta_n - \sum_{j=1}^{k} \delta_{ij} \right)} \tag{1}$$

$$\pi_{nix} = \frac{\exp\left(k\beta_n - \sum_{j=1}^{k} \delta_{ij} \right)}{1 + \sum_{k=1}^{m_i} \exp\left(k\beta_n - \sum_{j=1}^{k} \delta_{ij} \right)} \tag{2}$$

$$(x = 1,2, \ldots, m_i)$$

where δ_{ij} is a parameter associated with the transition between scores of $j-1$ and j on item i. The role of these item parameters is illustrated below. When Equations 1 and 2 are used to calibrate a set of polytomously-scored items, m_i parameters are estimated for each item i. One parameter β_n is estimated for each person n. This model is described by Masters (1982) and Masters and Wright (1984). Applications of the model are described by Wright and Masters (1982), Andrich (1982), Masters (1984), Adams (1985), and Koch and Dodd (1985, 1986).

Starting a Bank

An item bank can be established by gathering together a small set of items and administering them to a group of persons. These items are then calibrated and used as a core around which a bank can be developed. To illustrate a procedure for starting a bank, 14 items from the Watson-Glaser Critical Thinking Appraisal (Form A) were administered to a group of 368 Australian university students. Each of these 14 "items" is actually a question cluster, with all questions in the cluster referring to the same piece of introductory text (as in Figure 1). The 14 items vary in size, the smallest being

a cluster of two questions, the largest a cluster of six questions.

In general, a bank will contain dozens, and possibly hundreds, of items. This paper is not concerned with a bank of this size; rather, its purpose is to illustrate a procedure for calibrating a small core of items and then adding new items to this core. This procedure for adding new items can be repeated to build up a bank of any size.

To calibrate these initial 14 items, a count was made of the number of questions in each cluster answered correctly by each student. This gave each student a score between 0 (no questions in that cluster correct) and m_i (all m_i questions correct) on each item i. It should be noted that these scores do not indicate the specific questions a student answered correctly: A score of 2 on item i simply indicates that the student succeeded on two of the m_i questions in that cluster.

The resulting data were analyzed using CREDIT2[1], a microcomputer program for the Rasch analysis of ordered response categories. The estimation procedure implemented by this program is described in detail by Wright and Masters (1982) and is a generalization of the UCON procedure described by Wright and Stone (1979) for the dichotomous Rasch model. The results of the item analysis are shown in Table 1. The analysis provides a set of parameter estimates for each item, a standard error for each estimate, and a statistic summarizing the fit of each item to the model.

For an item containing three questions, CREDIT2 provides three estimates. But these estimates cannot be interpreted as the difficulties of the three questions in that item: The details of responses to individual questions within an item have not been retained in this analysis. Rather, the estimates for each item correspond to the transitions between the four response categories defined for that item (0—none of the three questions correct; 1—any one question in the cluster correct; 2—any two questions correct; 3—all three questions correct).

For example, Item 8 (CL08) is a three-question cluster for which the estimates −.95, −.01, and 1.45 logits were obtained. Students with ability estimates less than −.95 logits are estimated to be most likely to fail all three questions in this cluster to obtain a score of 0 on Item 8. For students with estimates between −.95 and −.01 logits, the most

[1]CREDIT2 is based on the program CREDIT (Masters, Wright, & Ludlow, 1981) and is available in both FORTRAN and BASIC from the authors.

Table 1
Parameter Estimates and Their Standard Errors for Each of 14 Item Clusters from Form A of the Watson-Glaser Critical Thinking Appraisal, and Item Fit Index for Each Item, for Group A

Item	Estimates						Errors						Fit
CL01	-1.41	-.14	.59	1.51	2.16		.42	.18	.13	.13	.16		1.51
CL02	-1.20	-.80	-.03	.72	1.41		.60	.29	.17	.13	.13		-1.16
CL03	-1.59	-.29	.36	.96	2.14	3.09	.51	.21	.15	.13	.15	.24	2.17
CL04	-1.25	-.80	-.57				.59	.26	.14				-.34
CL05	.14	-.94	-.08				.30	.22	.13				-1.25
CL06	-1.17	-.83	-.10				.52	.24	.13				-.70
CL07	-.03	-1.14	-.12	.48			.41	.29	.16	.12			-.76
CL08	-.95	-.01	1.45				.29	.14	.12				-.22
CL09	-2.40	-.10	-.12				.72	.19	.13				-.79
CL10	.53	.69					.14	.12					.80
CL11	-.30	-.56	.65				.28	.18	.12				.35
CL12	-1.52	-.11					.34	.13					-1.25
CL13	-.80	.93	2.54				.21	.12	.16				2.97
CL14	-1.04	-.66	.71				.40	.19	.12				-.75

probable outcome on Item 8 is estimated to be success on one (any one) of these three questions and failure on the other two. Between −.01 and 1.45 logits, the most probable outcome is success on any two questions in this cluster but failure on the other one. Students with ability estimates greater than 1.45 logits are estimated to be most likely to succeed on all three questions to make a score of 3 on Item 8.

This can be seen in Figure 2b. The $x = 0$ curve for Item 8 shows how the probability of failure on all three questions in Item 8 is modeled to decrease with increasing ability. The $x = 1$ curve shows how the probability of succeeding on only one of these three questions is modeled to increase and then decrease with ability, and so on. The three estimates (−.95, −.01, and 1.45 logits) for Item 8 correspond to the intersections of probability curves 0 and 1, 1 and 2, and 2 and 3.

Although Items 4, 8, and 9 in Figure 2 are all three-question clusters (and so are all scored from 0 to 3), the model probability curves for these three items are not identical. Differences between these three sets of probability curves reflect differences in the way the questions within these three items operate.

When ordered response categories are defined as counts of correct answers to a set of questions, regardless of which questions within that cluster were answered correctly, it is possible to ask what the shape of the model probability curves would be if all questions within that cluster were locally independent and of the same difficulty. The answer is that the curves would look something like the picture for Item 8: The pattern would be symmetrical with a spacing of 1.1 logits between successive intersection points (see Masters & Wright, 1984). If questions within a cluster are locally independent but *not* of the same difficulty, then the pattern of probability curves for that cluster becomes more dispersed.

The small spacing of the estimates for Item 4 (top of Figure 2) indicates that the three questions in this cluster are not functioning as locally independent items. There is a tendency for students to either succeed on all three questions in this cluster to obtain a score of 3, or to fail all three questions and score 0. The possibility of such dependence was, of course, a reason for not treating the three questions in Item 4 as independent dichotomously-scored items.

The pattern of probability curves for Item 9 at the bottom of Figure 2 suggests that one of the three questions in this cluster is very much easier than the other two. This is because there is a wide range of ability in which success on one question but failure on the other two is the most probable result. It is not possible to tell from this analysis which of the questions in Item 9 is easier than the other two. Notice also that for Item 9 there is no region of the ability continuum in which success on only two questions is the most probable outcome on this item: Respondents are most likely to succeed on only one question in Item 9 or to succeed on all three. (This can also be seen from the estimates −2.40, −.10, and −.12 for Item 9 which indicate that probability curves 2 and 3 intersect slightly to the left of the intersection of curves 1 and 2.) Again, the pattern of probability curves suggests a lack of independence among the questions in this cluster.

The statistic in the right-hand column of Table 1 summarizes the fit of each item to this model. This statistic (a standardized weighted mean square) was described by Wright and Masters (1982). When data conform to the model used here, this fit statistic has an expected mean of about zero, and an expected standard deviation near 1, meaning that the probability of this statistic exceeding +2 is about .02.

Two items (CL03 and CL13) in Table 1 have fit statistics greater than +2 and thus show relatively poor fit to the model. These two items appear not to be working in the same way as the other 12 critical thinking items. There may be some general problem with these two items, or the problem may be specific to this group of Australian students. It could, for example, be the result of an interaction of these items with cultural factors. Whatever the reason, before including Items 3 and 13 in a bank,

Figure 2
Model Probability Curves for Items 4, 8, and 9

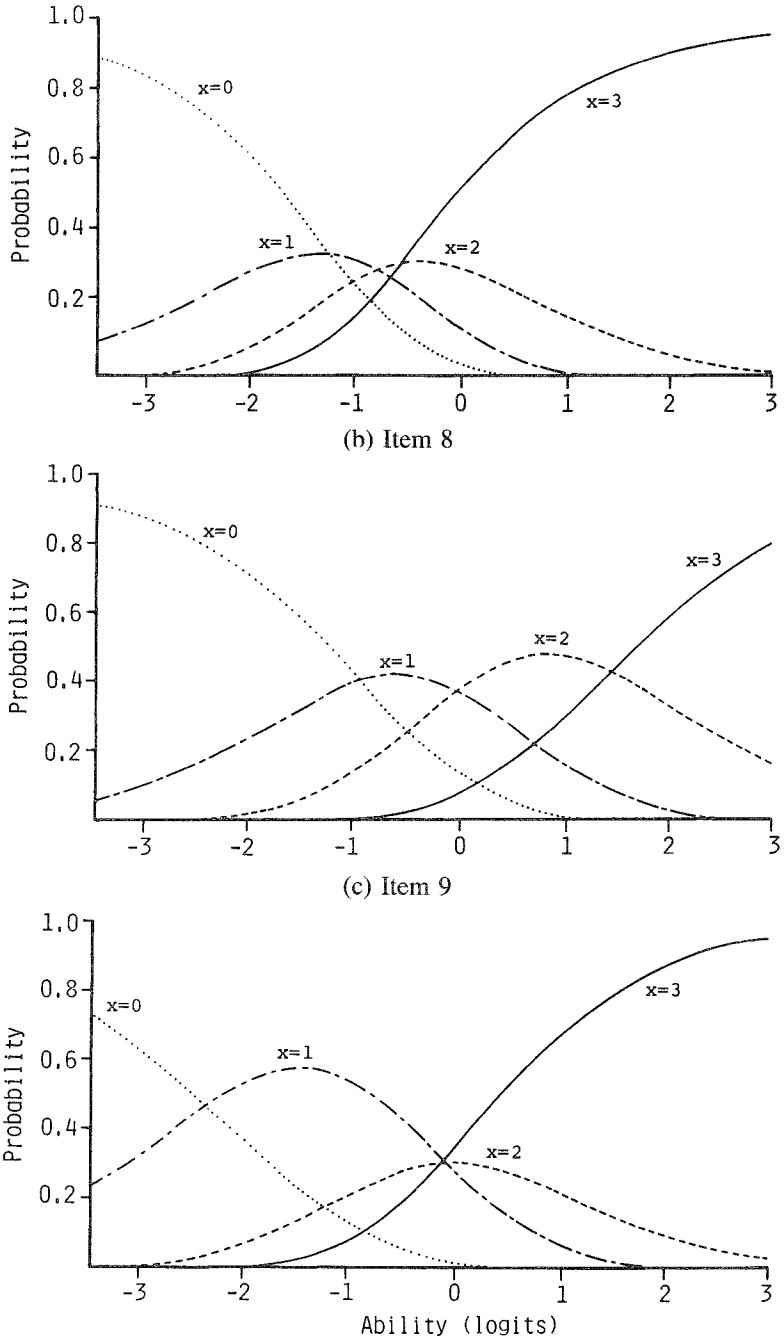
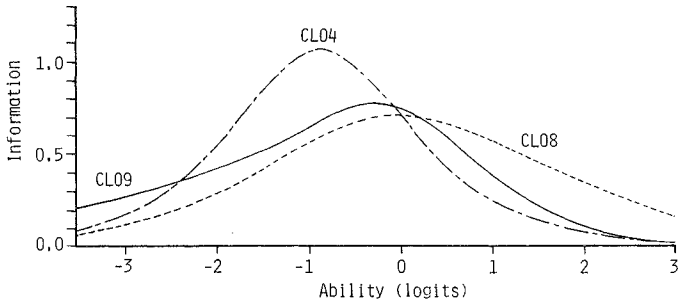


Figure 3
Item Information Curves for Items 4, 8, and 9



an investigation into the reasons for their behavior may be warranted.

Adaptive Testing

Once items have been calibrated, the estimates for each item can be attached to that item and stored in a bank. Items can then be selected from the bank to construct tests tailored to the abilities of individuals. In a computer adaptive test, items usually are selected to maximize the precision of each person's measure. This is done by selecting the item with the highest "information" level at the respondent's current ability estimate. (This is not the only possible decision rule for selecting items in an adaptive test, but it is the most common and the most convenient.) Under the measurement model used here, the information available from item i at ability $\hat{\beta}_n$ is given by

$$I_{ni} = \sum_{k=1}^{m_i} (k^2 P_{nik}) - \left(\sum_{k=1}^{m_i} k P_{nik} \right)^2 , \qquad (3)$$

where P_{nik} $(k = 1,2, \ldots, m_i)$ is the model probability of person n with estimated ability $\hat{\beta}_n$ succeeding on k questions in cluster i.

Figure 3 shows how the information I_{ni} available from Items 4, 8, and 9 (from Figure 2) varies with ability. These three items are differentially informative in different parts of the ability continuum. Item 8 is more informative (i.e., provides more precise measurement) than Items 4 and 9 for persons with ability estimates greater than about .5 logits. Item 4 is most informative between about -2 and 0 logits, and Item 9 is more informative than Items 4 and 8 only at very low levels of ability below -2 logits. This is due to the one very easy question in this cluster.

Computer adaptive testing with non-dichotomous items proceeds in exactly the same way as for dichotomous items (Weiss, 1982). During an adaptive test, a decision about which item to administer next is made by comparing levels of information for the unused items in the bank at the respondent's current ability estimate. Levels of item information might be calculated during a test, or they might be read from a table constructed prior to testing.

Adding to the Bank

Most item banks are dynamic. New items are added as they are developed and old items are discarded as they become irrelevant. The simplest way to incorporate new items into a bank is to administer them with some existing bank items. This provides a link between new items and the current bank and allows all items to be calibrated on the same bank scale.

To illustrate this linking procedure, 10 new critical thinking items of the same type as the 14 cluster items already in this bank were administered to a second group of 367 students, together with six items (CL09 to CL14) from the bank. The responses of this second group of students ("Group B") to

these 16 items were analyzed, and the results are shown in Table 2.

Two of the 16 items in Table 2 have relatively large positive fit statistics. The first of these, Item 13 with a fit value of 1.91, was one of the two items that showed evidence of misfit when it was calibrated on the first group of students (''Group A''). Its large fit value provides further evidence that it is not defining the same variable as the rest of these critical thinking items. The other misfitting item, Item 24, has an unusually large fit value of 4.43. This definitely warrants inspection.

Item 24 is a four-question cluster about the release of public school children during school time to attend religious instruction in their own churches. There are several plausible explanations for the aberrant behavior of this item. First, the use of the term ''public school'' can be expected to produce confusion. In Australia, a public school usually is one of a small number of elite church-owned schools outside the state system. The issue of religious instruction in these schools might be perceived differently from the issue of religious instruction in state schools. Second, one of the questions in this cluster describes religious instruction during school hours as being in violation of the Australian constitution. It is usual in Australian state schools for students to be released once a week during school hours to attend instruction in their own religions. This, supposedly, is not in violation of the Australian constitution. Third, to make a perfect score on this item, Australian students must describe the two arguments in favor of current Australian practice as weak and the two arguments against current practice as strong. Exactly the opposite is true of the American respondents on whom this item was normed. These cultural interactions almost certainly account for the very poor fit of Item 24 and suggest that this item should not be included in an Australian bank of critical thinking items.

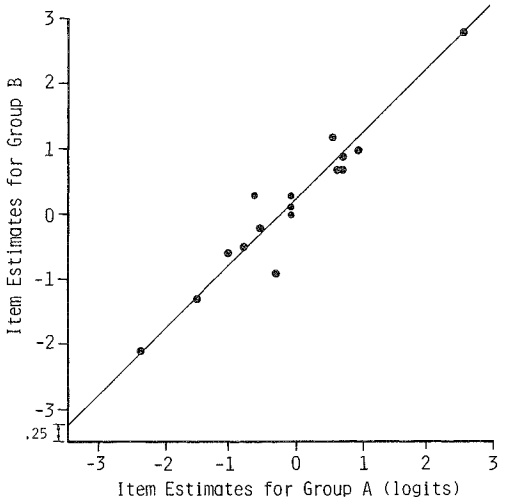
For each of the six link items (CL09 to CL14) taken by students in both Group A and Group B, two sets of estimates are now available. These estimates are expressed on different scales—the result of setting the mean of the 48 estimates in Table 1 to zero, and then setting the mean of the 47 estimates in Table 2 to zero. To bring all estimates to the same scale, an adjustment must be made for this difference in scale origin.

In Figure 4, the 16 estimates for these six link

Table 2
Parameter Estimates and Their Standard Errors for Each of 16 Item Clusters from Form A of the Watson-Glaser Critical Thinking Appraisal, and Item Fit Index for Each Item, for Group B

Item	Estimates				Errors				Fit
CL09	-2.09	.07	.28		.71	.19	.12		-.81
CL10	1.18	.69			.13	.12			.68
CL11	-.95	-.21	.69		.42	.18	.12		.34
CL12	-1.28	.01			.36	.13			-.60
CL13	-.49	.96	2.81		.21	.12	.16		1.91
CL14	-.61	.31	.88		.30	.15	.12		-1.83
CL15	-1.17	.02			.34	.13			-.90
CL16	.16	.33	.81		.22	.15	.12		-.95
CL17	-1.11	.30	.48		.39	.16	.12		-.97
CL18	.06	-.74	-.82	.62	.59	.40	.22	.12	.25
CL19	-1.07	-.36	.89		.46	.19	.12		-.66
CL20	-1.11	.31	.89		.37	.15	.12		-.22
CL21	-.66	-1.04	.74		.51	.24	.12		-.02
CL22	-1.07	.22	.98		.37	.15	.12		.66
CL23	-2.09	-.32	1.28		.72	.18	.12		-2.42
CL24	-.52	-.34	.76	1.30	.39	.21	.13	.12	4.43

Figure 4
Estimates for Link Items Calibrated on Student Groups A and B



items resulting from their calibration on Group B (Table 2) are plotted against their estimates from Group A (Table 1). The mean of the 16 estimates resulting from the calibration of these items on Group A and the mean of the 16 estimates resulting from their calibration on Group B were calculated, and a line with slope 1.0 is drawn through these two means. The intercept (.25 logits) on the Group B axis is the difference between the scale origins, and is therefore the amount that must be subtracted from the estimates in Table 2 to bring them to the bank scale. This simple linking procedure allows any new set of items to be added to a bank, provided that they first approximate the measurement model.

A final quality control check in this banking procedure is to compare the observed spread of points about the diagonal line in Figure 4 with their modeled variation. This check on the internal consistency of the common-item link is a check on the invariant operation of these six link items across the two student groups. For each of the points $(\hat{\delta}_{ijA}, \hat{\delta}_{ijB})$ plotted in Figure 4, a difference $\hat{\delta}_{ijA} - \hat{\delta}_{ijB}$ can be calculated and standardized to

$$Z_{ij} = \frac{\hat{\delta}_{ijA} - \hat{\delta}_{ijB}}{(S^2_{ijA} + S^2_{ijB})^{1/2}} , \qquad (4)$$

where S_{ijA} is the standard error of estimate $\hat{\delta}_{ijA}$, and S_{ijB} is the standard error of $\hat{\delta}_{ijB}$.

The inspection of these standardized differences for the 16 points in Figure 4 shows that almost all have values between -1 and $+1$; two are outside the range -2 to $+2$. When estimates from two calibration groups are very different, this indicates that one or more link items is not functioning in the same way in the two groups. It may be desirable to remove particularly erratic items from a link to improve the equating (see Wright & Bell, 1984). For these data, there is probably very little to be gained by not using all available link items.

Using the Bank

The application of this item banking procedure has resulted in a small bank of calibrated items. In view of its very poor fit, Item 24 should probably be removed from this bank, at least for Australian students. The anomalous behavior of this item lowers the utility of the bank for making item-free measures of critical thinking. Ideally, the second set of items (calibrated on Group B) should be recalibrated with Item 24 removed. The linking procedure might then be carried out again on the new set of estimates.

The item fit analysis has also raised a question about the validity of Items 3 and 13. Some further investigation of these two items is desirable. At this stage, Items 3 and 13 could probably be retained in the bank until some explanation for their less than ideal fit has been found.

With these items calibrated, other critical thinking items might be added to the bank in the same way. Initially, these new bank items might come from other forms of the Critical Thinking Appraisal already in existence. All that is required to add items to a bank is to administer them along with some existing bank items. Once a large bank is available, this can be used as a source of items for the construction of new test forms. And, because all bank items are calibrated on the same scale, scores on these different forms can be converted to the bank scale and compared directly.

The procedure for converting scores on new test forms to measures in logits on the bank scale is

straightforward. A user begins by selecting L items from the bank. Each item i drawn from the bank is scored from 0 to m_i, and is accompanied by m_i parameter estimates $\hat{\delta}_{i1}, \hat{\delta}_{i2}, \ldots, \hat{\delta}_{im_i}$. The maximum score that can be made on the resulting L-item test is $T = \sum_{i=1}^{L} m_i$.

Each score R between 1 and $T-1$ on this test is now converted to an ability estimate B_R (in logits) on the bank scale. The starting point is to define an initial ability value $B_R = \ln[R/(T-R)]$ for each score R. This is substituted into the following cycle, which is repeated until improvements in B_R become insignificant.

1. For each score R, start cycle: $A_R = B_R$.
2. For each item i, calculate

$$Q_{XR} = \exp\left(XA_R - \sum_{j=1}^{X} \hat{\delta}_{ij}\right) \quad (X = 1,2, \ldots, m_i) \tag{5}$$

$$P_{XR} = Q_{XR} \Big/ \left(1 + \sum_{K=1}^{m_i} Q_{KR}\right) \quad (X = 1,2, \ldots, m_i) \tag{6}$$

$$Y_i = \sum_{K=1}^{m_i} KP_{KR} \tag{7}$$

$$Z_i = \sum_{K=1}^{m_i} K^2 P_{KR} \tag{8}$$

3. The improved estimate

$$B_R = A_R + \left(R - \sum_{i=1}^{L} Y_i\right) \Big/ \sum_{i=1}^{L} (Z_i - Y_i^2) \tag{9}$$

is obtained, ending the cycle.

At the end of the first cycle, the new value of B_R is used to begin another cycle. At the end of each subsequent cycle, the new value of B_R is compared with the estimate from the previous cycle (now A_R). If $|B_R - A_R| < .01$, the procedure is terminated and the current value of B_R is used as the ability estimate corresponding to a score of R on this L-item test. The measurement error associated with B_R can be estimated as

$$S_R = \left[\sum_{i=1}^{L} (Z_i - Y_i^2)\right]^{-1/2}. \tag{10}$$

The above expressions can also be used for the selection of items in computer adaptive tests. The information provided by each item i at ability estimate B_R is simply $I_{Ri} = Z_i - Y_i^2$. During an adaptive test, I_{Ri} might be calculated for each item i at the current ability estimate B_R and used to select the most informative unused item. Alternatively, values of the item information might be calculated for a range of abilities prior to testing and used to construct item orders which could be stored and referred to during an adaptive test.

Discussion

The psychometric method used to transform a collection of items into a coherent measuring system is perhaps the most important part of an item bank. Items themselves are transient and expendable: They can be interchanged and replaced as required. But the bank scale upon which items are calibrated and persons are measured, and the psychometric method used to support this scale, are more permanent features of an item bank. Without a supporting psychometric method, an item bank ceases to be a measuring system and reverts to a mere collection.

The banking procedure described and applied in this paper is based on a simple extension of Rasch's dichotomous measurement model to responses scored in more than two response categories. In fact, it is probably the simplest possible extension of the Rasch model in that it applies the dichotomous model to each pair of adjacent response alternatives. An advantage of this simple formulation is that it permits the separation of person and item parameters during estimation. This makes conditional and unconditional maximum likelihood estimation straightforward (see Wright & Masters, 1982), and makes the item banking procedure described in this paper simple enough to be applied routinely on a microcomputer.

The application of this procedure allows item banks to be extended to incorporate a wider variety of item response formats than just correct/incorrect scoring. When this method is used, parameter estimates for an item will be valid from occasion to occasion only if the ordered performance levels in

that item are defined in more or less the same way each time the item is used. To ensure comparability, it may be necessary to provide bank users with descriptions of the criteria to be applied in rating performances on bank items or in awarding partial credit. In an educational context, objectivity might be further improved by including samples of student work as illustrations of scoring criteria. Provided that a few simple precautions of this type are taken, it should be possible for constructors of item banks to use the method described in this paper to calibrate a wider variety of item types and to become more adventurous in their choice of materials for inclusion in an item bank.

References

Adams, R. J. (1985). *Diagnostic adaptive testing with the Rasch partial credit model* [Working Paper No. 2]. Education Department of Victoria, Australia.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105–113.

Choppin, B. H. (1968). An item bank using sample-free calibration. *Nature, 219,* 870–872.

Choppin, B. H. (1976). Recent developments in item banking. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 233–245). New York: Wiley.

Choppin, B. H. (1978). Item banking and the monitoring of achievement. *Research in progress series* (No. 1). Slough, England: National Foundation for Educational Research.

Choppin, B. H. (1981). Educational measurement and the item bank model. In C. Lacey & D. Lawton (Eds.), *Issues in evaluation and accountability*. London: Methuen.

Cornish, G., & Wines, R. (1977). *Mathematics profile series*. Hawthorn, Victoria: Australian Council for Educational Research.

Elliott, C. D. (1983). *British Ability Scales, Manuals 1–4*. Windsor, England: NFER-Nelson.

Forster, F., & Ascher, G. (1977). *The Rasch calibrated item bank: A new tool for competency based evaluation*. Portland OR: Portland Public Schools, Oregon State Department of Education.

Hill, P. W. (1985). *The Tests of Reading Comprehension (TORCH)*. Paper presented at the annual meeting of the International Association of Educational Assessment, Oxford.

Koch, W. R., & Dodd, B. G. (1985). *Computerized adaptive attitude measurement*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koch, W. R., & Dodd, B. G. (1986). *Operational characteristics of adaptive testing procedures using partial credit scoring*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Koslin, B., Koslin, S., Zeno, S., & Wainer, H. (1977). *The validity and reliability of the Degrees of Reading Power test*. Elmsford NY: Touchstone Applied Science Associates.

Kubinger, K. D. (1985, August). *On a Rasch model based test for noncomputerized adaptive testing*. Paper presented at the 13th IPN (Institut für die Pädagogik der Naturwissenschaften) conference on latent trait and latent class models in educational research, Kiel, Federal Republic of Germany.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G. N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement, 21,* 19–32.

Masters, G. N., & Adams, R. J. (1985, March). *A latent trait method for interactive achievement tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

Masters, G. N., Wright, B. D., & Ludlow, L. H. (1981). *CREDIT: A Rasch program for ordered response categories*. Chicago: MESA Psychometrics Laboratory, University of Chicago.

Pollitt, A. B. (1979). Item banking. *Issues in educational assessment*. Edinburgh: Scottish Education Department.

Pollitt, A. B. (1985). Item banking and school assessment. In N. Entwistle (Ed.), *New directions in educational psychology*. East Sussex, England: The Falmer Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14,* 58–94.

Tognolini, J. (1982). *Pupil achievement in stage 6 mathematics* [Discussion paper No. 15]. Perth: Education Department of Western Australia.

Trismen, D. M. (1981). *The development and administration of a set of mathematics items with hints* (Re-

search Bulletin 81-5). Princeton NJ: Educational Testing Service.

Trismen, D. M. (1982). *Mathematics items with hints* (Research Bulletin 82-11). Princeton NJ: Educational Testing Service.

Trismen, D. M. (1983). *Microcomputer administration of mathematics items with hints* [unpublished report]. Princeton NJ: Educational Testing Service.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Whetton, C., & Childs, R. (1981). *The effects of item-by-item feedback given during an ability test*. Unpublished report, National Foundation for Educational Research, England.

Wilcox, R. R. (1982). Some new results on an answer-until-correct scoring procedure. *Journal of Educational Measurement, 19,* 67–74.

Wongbundhit, Y. (1985). *Item banking procedure and quality control in Dade County public schools*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21,* 331–345.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Acknowledgments

The authors thank Roger Arwas of the Victorian Public Service Board for making these data available.

Author's Address

Send requests for reprints or further information to G. N. Masters, Centre for the Study of Higher Education, University of Melbourne, Parkville, Victoria 3052, Australia.