Effects of Variations in Item Step Values on Item and Test Information in the Partial Credit Model

Barbara G. Dodd and William R. Koch
University of Texas at Austin

Simulated data were used to investigate systematically the impact of various orderings of step difficulties on the distribution of item information for the partial credit model. It was found that the distribution of information for an item was a function of (1) the range of the step difficulty values, (2) the number of step difficulties that were out of sequential order, and (3) the distance between the step values that were out of order. Also, by using relative efficiency comparisons, the relationship between the step estimates and the distribution of item information was used to demonstrate the effects of various test revisions (through the addition and/or deletion of items with specific step characteristics) on the resulting test's precision of measurement. The usefulness of item and test information functions for specific measurement applications of the partial credit model is also discussed.

During the last decade, developments in item response theory (IRT) have offered new approaches for solving many practical measurement problems. Birnbaum's (1968) conceptualization of information functions for individual items and tests has been used in many applications of IRT. The primary benefit of information functions is that they allow selection of items for inclusion in a test such that the precision of measurement for the test is maximized at the specific trait (θ) level that is of interest to the examiner. Another benefit is that information functions for two tests can be compared in terms of relative efficiency, which can aid in the selection of the best test for a given measurement situation. Information functions have also been used effectively to determine item selection for computerized adaptive testing.

For dichotomously scored items, information functions have primarily been used with the three-parameter IRT model rather than the one-parameter Rasch model. The information an item provides is by definition the square of the ratio of the slope of the item characteristic curve to the conditional standard error of measurement (Lord, 1980). For the three-parameter model, item information functions for items in a test differ from one another because items are free to vary in terms of discriminations, difficulties, and the lower asymptotes of the item characteristic curves. For the Rasch model, all item information functions yield the same maximum amount of information; they differ from one another only in terms of the θ level for which the maximum information is provided. This is because the Rasch model assumes that items have equal discriminations and lower asymptotes of 0 for the item characteristic curves. Thus the use of information functions with the simple Rasch model usually provides no additional information beyond the difficulty level of the items.

Samejima (1969) extended Birnbaum's formulation of information functions to the case where items are polychotomously scored. By comparing the information yielded by items scored with op-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 4, December 1987, pp. 371–384

0146-6216/87/040371-14$1.95

timal dichotomization with the information yielded by scoring the items according to the graded response model (ordered categories), Samejima (1969, 1976) found that the graded response approach yielded considerably greater precision of measurement. Koch (1983) also used graded response information analyses of a Likert-type attitude scale to demonstrate some advantages of the graded response model over traditional Likert scaling procedures. Using Samejima's formulation of information functions for the polychotomous case, Bock (1972) and Thissen (1976) demonstrated that the nominal model (in which categories are not assumed to be ordered) provides more information, particularly for lower levels of the trait continuum, than dichotomous scoring of the same items.

Dodd and Koch (1985) applied Samejima's formulation of information functions for polychotomously scored items to the partial credit model, which is a member of the Rasch family of models. Unlike the simple Rasch model for dichotomously scored items, it was found that item information functions for the partial credit model could differ substantially from one another as a function of the step estimates for each item. The different shapes of the information functions for items were a function of the proximity of the first and last step estimates. Items with a small distance between the first and last step estimates provided the most information, but only for a narrow range of the trait continuum. Items with a large difference between the first and last step estimates had less peaked information functions, but information was provided for a wider range of the trait continuum. It was also found that in all cases, an item provides maximum information for the trait levels that are in the range of the step estimates.

This paper presents the results of a further investigation of the relationship between the distribution of information for an item and the item step estimates. Also, by means of relative efficiency comparisons, the relationship between the step estimates and the distribution of item information was used to demonstrate the effects of various test revisions on the resulting test's precision of measurement.

The Partial Credit Model

The partial credit model (Masters, 1982) is a model in the Rasch family developed specifically for the case of polychotomously scored items. Like the graded response model, it is appropriate when responses to an item can be evaluated according to the degree of attainment of a solution to a problem or the magnitude of agreement with an attitude statement. That is, responses to item i are classified into $(m_i + 1)$ ordered categories so that lower-numbered categories represent less of the latent trait measured by the item than do higher-numbered categories. The category scores for item i are successive integers, denoted x_i, that can take on the values $0, 1, \ldots, m_i$.

Masters (1981, 1982) interpreted the ordered category scores for an item to represent the number of subtasks or steps in an item that had been successfully completed. The step interpretation is particularly appropriate for items on which partial credit can be awarded for a partially correct solution to a problem. For instance, in the following three-step mathematical problem,

$$[(5.5/.5) - 2]^3 \quad ,$$

the first subtask or step is to solve the division $5.5/.5 = ?$, the second step is to solve the subtraction $11 - 2 = ?$, and the third step is to solve the cube $9^3 = ?$. A correct solution to only the first step will result in a category score of 1, because a single step has been successfully completed. A correct answer to the second step is impossible (except by chance) without a correct solution to the first step. Consequently, a correct solution to the second step will result in a category score of 2, because two steps have been successfully completed. The maximum score of 3 will be awarded for a correct solution to the third step, because it requires the successful completion of the prior two steps.

The step interpretation is not restricted to items on which partial credit can be awarded for a partially correct solution to a problem; it is appropriate for any item for which the response alternatives can be ordered, assuming good fit of the model to the data. An example of an ordered-response item would be an attitude statement that has response

alternatives ranging from "strongly disagree" to "strongly agree".

It should be noted that successful completion of a given step depends on the successful completion of the prior step or steps. In fact, the partial credit model requires that steps within an item be taken in the same order by all individuals. This, however, does not in turn require that the difficulties of the steps be ordered so that later steps are more difficult than earlier steps. For example, in the mathematical item presented above, the second step, which involved subtraction, was easier than the first and third steps, which involved division and raising a number to a power, respectively. The partial credit model, therefore, requires only that the steps within an item be ordered in sequence, not that they be ordered in terms of difficulty.

For the partial credit model, the probability that an individual will respond in a given category can be written as

$$P_{xi}(\theta) = \frac{\exp\left[\sum_{j=0}^{x_i}(\theta - b_{xj})\right]}{\sum_{k=0}^{m_i}\exp\left[\sum_{j=0}^{k}(\theta - b_{xj})\right]} . \qquad (1)$$

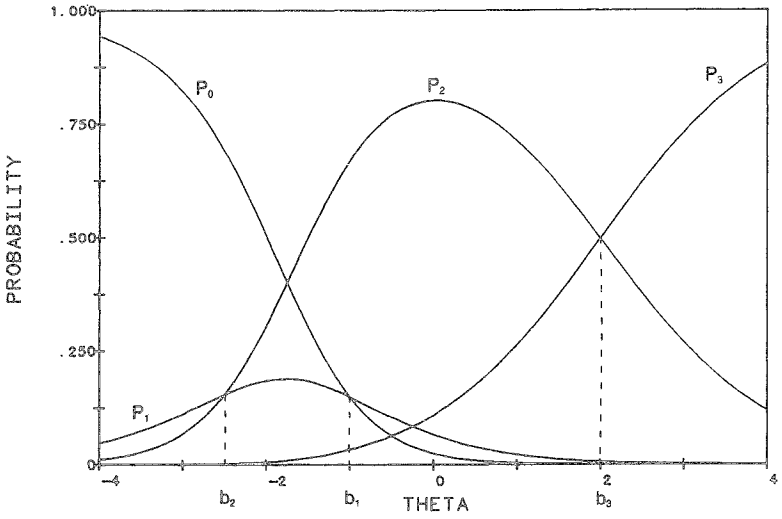
Equation 1 is the general form for obtaining the operating characteristic curve for category score x_i with the partial credit model. The θ term is the trait level, and the b_{xj} $(x_j = 1, \ldots, m_i)$ term is the difficulty of the step associated with the category score x_i. For notational convenience, Masters (1982) defined the $\sum_j(\theta - b_{xj})$ for $j = 0$ to 0 as being equal to zero.

Equation 1 was applied to the hypothetical mathematical item used above to illustrate the step interpretation; the resulting operating characteristic curves are presented in Figure 1. The step difficulty parameters, b_{xj}, correspond to the points of intersection of adjacent operating characteristic curves.

Information Functions for the Partial Credit Model

Dodd (1984/1985) applied Samejima's (1969) formulation of information functions for polychotomously scored items to the partial credit model. Because Samejima's equation for item information was expressed in terms of the operating characteristic curves for an item, the equation was simply applied to the partial credit model's operating characteristic curves for an item. Samejima defined the

Figure 1
Operating Characteristic Curves for a Four-Category Partial Credit Item



item information function as

$$I_i(\theta) = \sum_{x_i=0}^{m_i} \frac{[P'_{x_i}(\theta)]^2}{P_{x_i}(\theta)} - \sum_{x_i=0}^{m_i} P''_{x_i}(\theta) \quad , \tag{2}$$

where $P_{x_i}(\theta)$ is the probability of receiving category score x_i conditional on θ, and $P'_{x_i}(\theta)$ and $P''_{x_i}(\theta)$ are the first and second derivatives of $P_{x_i}(\theta)$, respectively. The second term of Equation 2 is equal to 0.0 and thus can be deleted from the equation for item functions.

The test information function is simply defined as the sum of the item information functions:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n_i} \sum_{x_i=0}^{m_i} \frac{[P'_{x_i}(\theta)]^2}{P_{x_i}(\theta)} \quad . \tag{3}$$

Thus, as is the case for dichotomously scored items, the information that a polychotomously scored item contributes to the test information function is independent of the information provided by the other items in the test.

Method

Datasets

Step difficulty values for the item information analyses. The relationship between the item information and the ordering of the step difficulty values for the partial credit model was assessed with four-step and three-step items. For the four-step items, the orderings of the step difficulty values of −1.0, −.5, .5, and 1.0 were systematically varied to yield 24 items that differed from one another only in terms of the ordering of the step difficulty values. All possible orderings of the three step difficulty values of −1.0, 0.0, and 1.0 yielded 6 three-step items.

Simulated data for the relative efficiency analyses. A simulated dataset was constructed to demonstrate the effects of various test revisions (through the addition and/or deletion of items with specific step characteristics) on the resulting test's precision of measurement. The decision to use a simulated dataset rather than calculating information directly from specified item parameters was based on the fact that in practice, researchers do not have advance knowledge of parameters. It is more realistic to use estimates of parameters (as usually obtained in a calibration program) to demonstrate the relationship between those estimates and the resulting distribution of information, and their consequent impact on the precision of measurement when various test revisions are made. The demonstration is the same whether known or estimated parameters are used, but the advantage of estimated parameters obtained from the calibration of a simulated dataset is that they are more representative of the values that might be obtained in practice.

The simulated dataset consisted of simulated responses to 240 items from 1,000 hypothetical examinees. The data were generated specifically to fit the partial credit model. Each item was constructed to have four response alternatives; thus, three step difficulty parameters were specified for each item in the generation program. The item parameters used to generate the data were deliberately specified to be similar to the types of item parameters obtained with real data. Although the item parameters used to generate the data were based on typical values obtained empirically with real data in previous research (Koch & Dodd, 1985; Masters, 1982, 1984), the step parameters were also systematically varied to demonstrate the effects of various patterns of step values on information and, in turn, the relative efficiency of various test revisions. A conscious effort was made to spread the items uniformly across the trait continuum so that one-third of the items were very easy (all negative step values), one-third were of moderate difficulty (a mixture of positive and negative step values), and one-third were very difficult (all positive step values). In addition, the item parameters for the items at each difficulty level were specified so that half of the items had step difficulty values that would yield peaked information functions, and the other half would yield item information functions that were relatively flat. Some of the items at each of the difficulty levels also had step difficulty values specified so that later steps would be easier than earlier steps.

The data generation procedure first selected a z value from a normal distribution (0,1) to represent the first person's θ level. The probability of the person receiving each category score for the first item was then calculated based on the known item parameters. The probabilities were then summed to obtain cumulative subtotals for each category score from 0 to 3. The subtotals for each category score served as boundaries between response alternatives. After the boundaries were calculated, a value was drawn randomly from a uniform distribution that ranged from 0.0 to 1.0. The random value was compared to the category score boundaries to determine the category score interval into which the random number fell.

The same z value and procedures were used to obtain the first person's category scores for the remaining items. After these responses were generated, a new z value was randomly selected to obtain the category scores for the second person. This procedure continued until category scores had been generated for 1,000 persons.

Calibration Procedures for the Simulated Data

The step difficulty parameters for the items and the person parameters for the simulated data were estimated by the PARTIAL computer program, which was written according to the calibration procedures described by Masters (1982) in his presentation of the partial credit model. PARTIAL performed maximum likelihood estimation and employed a Newton-Raphson iteration procedure to obtain the estimates of the respondents' trait levels and the step difficulties for the items. Iterative cycles were continued until all the item parameter estimates had converged. To ensure stable estimates of the item parameters, two separate calibration runs were conducted. The 120 items that were generated to yield peaked item information functions were calibrated in one run, while the 120 items that were generated to yield flat item information functions were calibrated in another run. The item parameter estimates from the two calibration runs were on the same scale of measurement because (1) the responses from the same 1,000 examinees were used for both calibrations, and (2) the PARTIAL program sets the origin of the θ scale at the mean of the trait estimates, with a variance of 1.0 set as the unit of measurement.

Information Analyses

The four step values (-1.0, $-.5$, .5, and 1.0) and three step values (-1.0, 0.0, and 1.0) that were used to investigate the effects of various orderings of the step values on item information were treated as known parameters in the information analyses. Equation 2 was used to calculate information for the θ values ranging from -4.0 to 4.0 at intervals of .1 for the 24 items representing all possible orderings of the four step difficulty parameters and the six items representing all possible orderings of the three step difficulty parameters.

Relative Efficiency Analyses

Item parameter estimates yielded by PARTIAL were used to demonstrate the effects of various test revisions through relative efficiency analyses. A base test of 60 items was constructed from the 240-item pool. Thirty of the items were selected from the 120 items that were generated to yield peaked information functions, while the remaining 30 items were selected from the 120 items that were generated to yield flat information functions. For each of the 30-item subsets of the base test, 10 items were easy, 10 items were of moderate difficulty, and 10 items were difficult.

Because the simulated data were generated according to prespecified step difficulty values, it was possible to investigate the effects of the range of the step estimates, the three levels of item difficulty, and various orderings of the step estimates on the precision of measurement of the resulting test. These effects were studied when items with certain step estimate characteristics were systematically added to or deleted from the base test. The impact of the various test revisions on the resulting test's precision of measurement (compared to the precision of measurement of the base test) was

assessed using relative efficiency analyses. Relative efficiency was defined as the ratio of the information of the revised test to the information of the base test for each trait level (Lord, 1980).

Results

Item Information Analyses

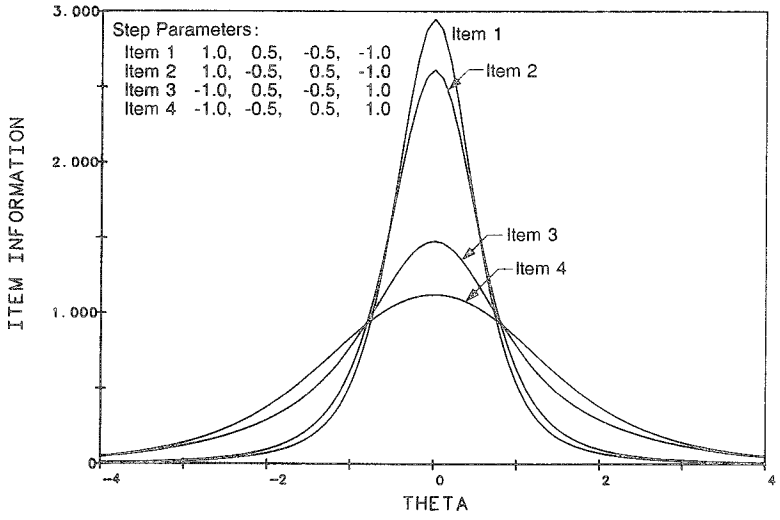
The item information functions for the 24 four-step items confirmed the finding of Dodd and Koch (1985) that across the entire trait continuum, all items provided the same total amount of information. That is, the areas under the item information curves were equal. While the 6 three-step items revealed the same phenomenon, it was discovered that the four-step items yielded more total information across the entire trait continuum than the three-step items. Thus, only items with the same number of steps yield the same total amount of information across the entire trait continuum.

Inspection of the step difficulty parameters revealed that the maximum information yielded by an item was directly related to (1) how close the first step difficulty parameter was to the last step difficulty parameter for the item, and (2) the number and magnitude of the reversals in the sequential order of the step difficulty parameters.

When the proximity of the first and last step difficulty parameters was held constant, the degree of peakedness of the item information function was related to the degree to which the step difficulty parameters deviated from sequential ordering. The 4 items with four-step parameters that had a distance between the first and last step parameters of 2.0 were selected to demonstrate this effect. Figure 2 presents the item information functions for these four items.

The flattest information function was for Item 4, which had the step estimates in sequential order from easiest to most difficult (-1.0, $-.5$, .5, and 1.0). The most peaked information function was for Item 1, which had step difficulty parameters in reverse order from most difficult to easiest (1.0, .5, $-.5$, and -1.0). Items 2 and 3 yielded maximum information values that were between the maximum information values for Items 1 and 4. It should be noted that the step difficulty parameters

Figure 2
Item Information Functions for Four Items That Have the Same Proximity of the First and Last Step Parameters but Differ in the Number of Step Parameters Out of Sequential Order



for Items 2 and 3 had some of the step parameters out of sequence, such that later steps were easier than earlier steps. For Item 2 the first and last step parameters were reversed, while for Item 3 the second and third step parameters were reversed. Item 2 yielded a more peaked information function than Item 3 because the step parameters that were reversed represented nonadjacent steps.

Collectively, these information function analyses revealed that when the distance between the first and last step parameters is held constant, the more that the step parameters are out of sequential order and the greater the number of positions the steps are displaced, the more peaked the information function will be. This finding also held for the three-step items.

When both the number of steps that were not out of sequence and the degree of displacement of the steps were held constant, items with a small distance between the first and last step parameters yielded more information than items with a larger distance between the first and last step parameters. This relationship is demonstrated with Items 5 and 6 in Figure 3.

Both Items 5 and 6 had a single step parameter that was out of sequential order by two steps. Item 6, which had a distance between the first and last step parameters of .5, yielded a slightly more peaked information function than Item 5, which had a distance between the first and last step parameters of 1.5. Thus, the closer the proximity of the first and last step parameters, the more peaked was the information function. This finding replicated the results of Dodd and Koch (1985).

It was also discovered that the size of the distance between the step parameters that were out of sequential order can override the effects of the proximity of the first and last step parameters on the peakedness of the information function. Figure 4 shows two items, each item having a reversal of adjacent step difficulty parameters. For Item 7, the second and third steps were reversed, and the distance between the first and last step parameters was 2.0. For Item 8, the third and fourth step parameters were reversed, and the distance between the first and last step parameters was 1.5. Item 7, which had the greater distance between the first and last step parameters, yielded a more peaked information function because the distance between the two reversed steps (1.0) was greater than the distance

Figure 3
Item Information Functions for Two Items That Have the Same Number of Step Parameters Out of Sequential Order but Differ in the Proximity of the First and Last Step Parameters

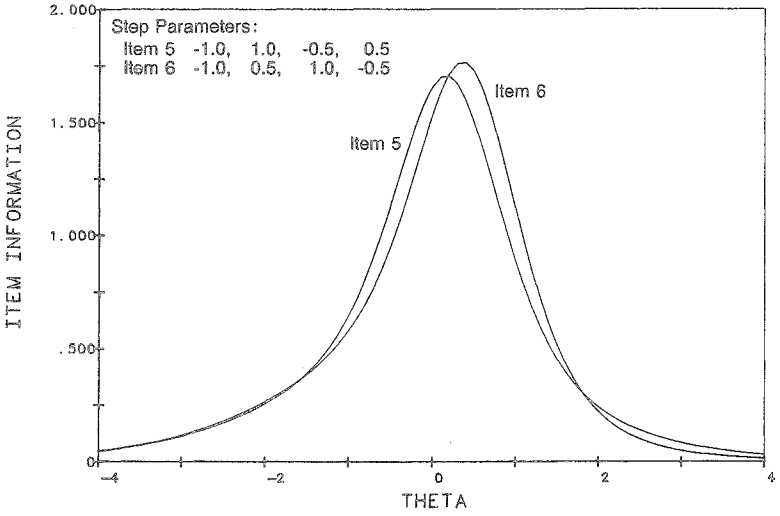
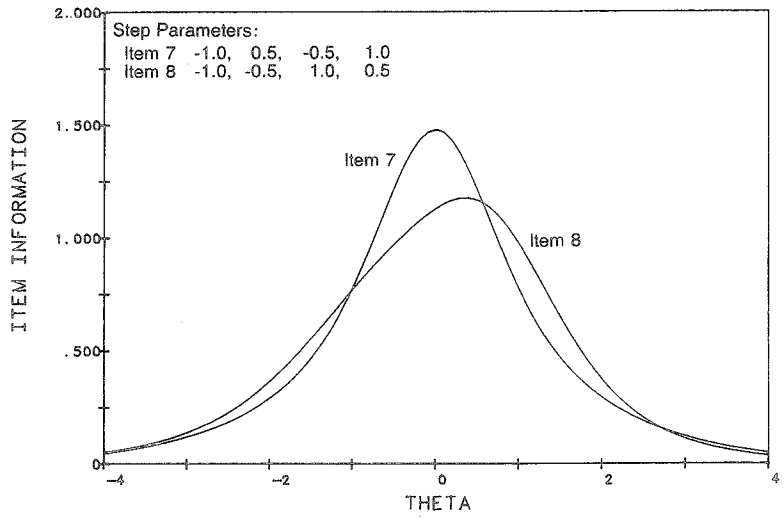


Figure 4
Item Information Functions for Two Items That Differ in the Proximity of the First and Last Parameters as Well as the Magnitude of the Distance Between the Reversed Step Parameters



between the two reversed steps for Item 8 (.5). The effect of the proximity of the first and last step parameters on the peakedness of the information function was dominated by the distance between the reversed step parameters. Thus, the greater the magnitude of the distance between the steps that were out of sequential order, the more peaked was the information function.

Collectively, the item information analyses revealed that the degree of peakedness of the item information function depends not only on the proximity of the first and last step parameters, but also on the number of steps that are out of sequential order and the magnitude of the distance between the reversed step parameters.

Relative Efficiency Analyses

To study the effects of various test revisions on the precision of measurement, it was necessary to categorize the 60 items in the base test and the remaining 180 items from the initial pool according to the range, level of difficulty, and ordering of the step estimates. Inspection of the item parameter estimates revealed two possible classifications for the range factor. If the range of the step estimates for an item was less than 1.5, the item was arbitrarily classified as a small-range item. If, however, the range of the step estimates was greater than 2.0, the item was classified as a large-range item. The items that met the criterion for inclusion in one of the range categories were then subdivided into three levels of difficulty. Items with negative step estimates were considered to be easy items, while items with positive step estimates were classified as difficult items. Those items with step estimates that centered around 0 were considered to be of moderate difficulty. Finally, the selected items were classified according to whether the step estimates were in sequential order.

After the base test items were classified according to the two levels of range (small, large), the three levels of item difficulty (easy, medium, difficult), and the two levels of ordering of the step estimates (sequential, nonsequential), 9 of the 12 possible conditions had five items that could be systematically deleted from the base test to determine the revised test's precision of measurement

relative to the base test. Three conditions had an insufficient number of items to assess the effects of deleting items from the base test: (1) large-range, easy items with nonsequential step estimates; (2) small-range, difficult items with sequential step estimates; and (3) large-range, difficult items with sequential step estimates.

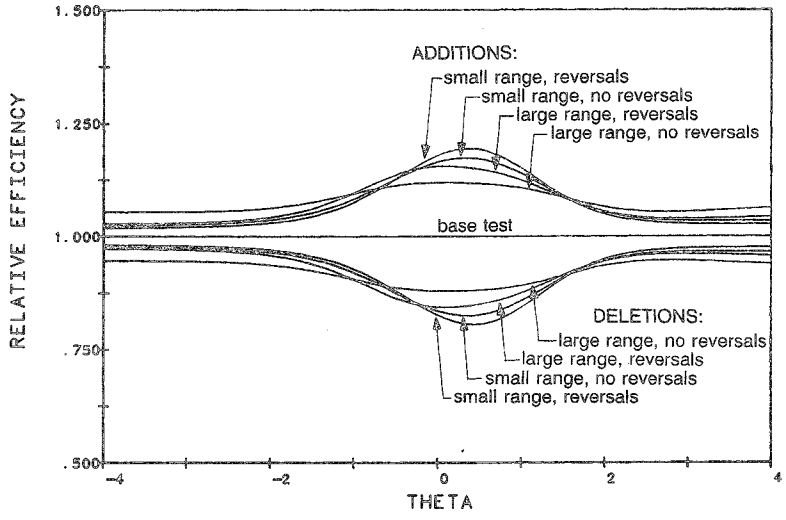
The three-way classification system of the 180 items that were not included in the base test yielded 8 conditions that had five items that could be systematically added to the base test to determine the effect on the resulting test's precision of measurement. Four conditions had an insufficient number of items to determine their effect on test revisions: (1) small-range, difficult items with step estimates in sequential order; (2) large-range, difficult items with step estimates in sequential order; (3) large-range, difficult items with step estimates that were not in sequential order; and (4) large-range, easy items with step estimates that were not in sequential order.

In order to facilitate the comparison of the precision of measurement yielded by each of the 17 test revisions relative to the precision of measurement of the base test, relative efficiency plots were constructed. Relative efficiency in this study was defined as the ratio of the information for the revised test to the information of the base test.

Figure 5 presents the relative efficiency comparisons for the 8 test revisions involving the items of moderate difficulty. The 4 test revisions that involved the deletion of five items of moderate difficulty lost most of their efficiency in the middle of the θ scale. The 4 test revisions that involved the addition of five medium-difficulty items were more efficient, particularly in the middle of the θ scale, relative to the base test.

The effects of the addition or deletion of items with certain step characteristics on the precision of measurement of the revised test relative to the base test were in the expected direction. Items with step estimates that covered a small range and were not in sequential order produced the largest change in the revised test's efficiency relative to the base test in the middle of the θ scale. The next largest change in efficiency was for the revisions that involved the items with step estimates that spanned a small range and were in sequential order. The revisions that included the addition or deletion of items with step estimates that covered a large range and were out of sequential order showed the next largest change in efficiency relative to the base test. The

Figure 5
Relative Efficiency of the Eight Test Revisions That Involved the Addition or Deletion of Medium-Difficulty Items



revised tests that yielded the smallest change in efficiency relative to the base test were the tests that involved the addition or deletion of items with step estimates that spanned a large range and were in sequential order.

The relative efficiency comparisons for the six test revisions that involved the addition or deletion of easy items are depicted in Figure 6. For θ values of −.6 to −2.4, the items with certain step characteristics that were deleted from or added to the base test produced the same ordering of the magnitude of the change in efficiency, relative to the base test, that was observed with the test revisions involving the items of medium difficulty. That is, the items with step estimates that covered a small range and were out of sequential order produced the largest change in the efficiency of the revised test relative to the base test. The next largest change in efficiency was for the revised tests that had small-range items with step estimates with no reversals added to or deleted from the base test; these tests were followed by test revisions that involved items with step estimates that spanned a large range and were in sequential order. It should be noted, however, that this ordering of the magnitude of change in the efficiency of the revised tests was reversed for θ levels less than −2.4. The reason for this result was that large-range items yielded information over a wider range of the θ scale than did small-range items.

Figure 7 presents the relative efficiency analyses for the three test revisions that involved the addition or deletion of the difficult items. These results mirror those found for the test revisions with the items of moderate difficulty and the easy items. The largest changes in the efficiency of the revised tests occurred within the range of the step estimates.

The effects of revising a test to maximize the precision of measurement for a given θ level for a given measurement situation were also investigated. The base test was revised so as to maximize the precision of measurement at the upper end of the θ scale. This would be a desirable test revision if the test constructor wished to increase the test's ability to discriminate among high-ability individuals who, for example, might be competing for a scholarship. To demonstrate the effect of this type of revision, the 20 easy items and the 20 items of moderate difficulty were first deleted from the base test. Thus only the 20 difficult items remained in

Figure 6
Relative Efficiency of the Six Test Revisions That Involved the Addition or Deletion of Easy Items

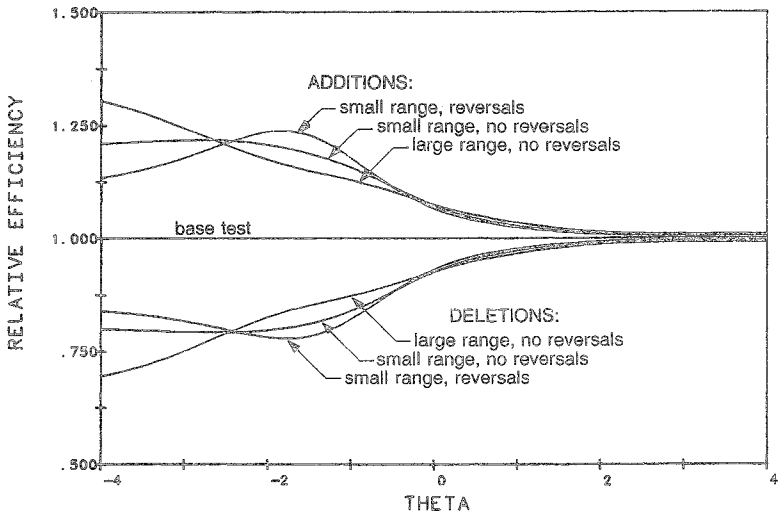
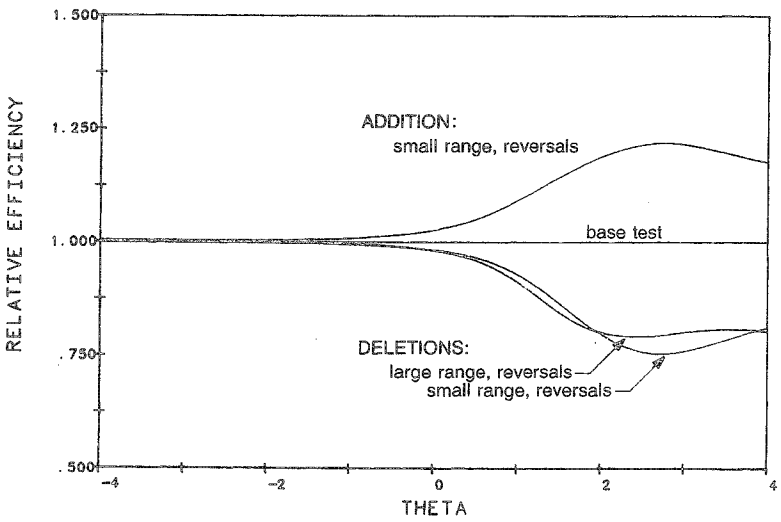


Figure 7
Relative Efficiency of the Three Test Revisions That Involved the Addition or Deletion of Difficult Items



the base test. Next, the 40 difficult items that yielded the most information at a θ level of 2.5 were selected from the 60 difficult items in the item pool that were not included in the base test. These additional 40 difficult items were added to the base test and a relative efficiency analysis was conducted. The result of this analysis is presented in Figure 8.

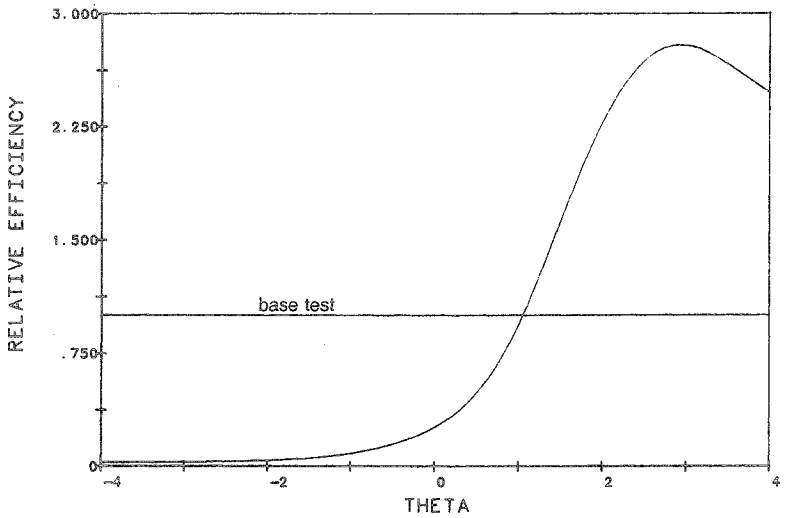
As expected, relative to the base test, the revised test was considerably less efficient for θ levels below 1.0. This is because there were no easy items or items of moderate difficulty included in the revised test. More interesting was the fact that the revised test was almost three times as efficient as the base test at the upper end of the θ scale. Thus, the test revision accomplished the task of improving the precision of measurement for high-ability examinees.

Discussion

The item information analyses confirmed some of the findings of Dodd and Koch (1985) and provided further clarification and extension of other findings. Both studies demonstrated that the maximum information provided by an item calibrated according to the partial credit model was provided within the range of the step difficulty values. It was also found that, across the entire trait continuum, items with the same number of score categories provided the same total amount of information. The finding that items with more score categories yielded more total information across the entire θ scale than items with fewer score categories was not surprising. This finding is consistent with the belief that items that are more heavily weighted (i.e., worth more points) provide more information or allow for finer discriminations among examinees than items that are worth fewer points.

The systematic comparisons of item information functions in this study revealed that the peakedness of the item information functions was not solely due to the proximity of the first and last step estimates, as was found in previous research. Rather, it was discovered that the number of reversals and the magnitude of the distance between the steps that were out of sequential order could dominate the effects of the proximity of first and last step estimates on the peakedness of the information function. That is, the greater the number of step

Figure 8
Relative Efficiency of the Test Revision That Replaced the 20 Easy and 20 Medium-Difficulty Items With 40 Difficult Items



estimates out of sequential order and the greater the magnitude of the distance between the step estimates that were out of order, the more peaked was the information function. The most peaked information functions were found for items with step values that were ordered from most difficult to easiest.

It is appropriate to question whether an item with a completely reversed sequential order of the step estimates is possible. A mathematical item such as

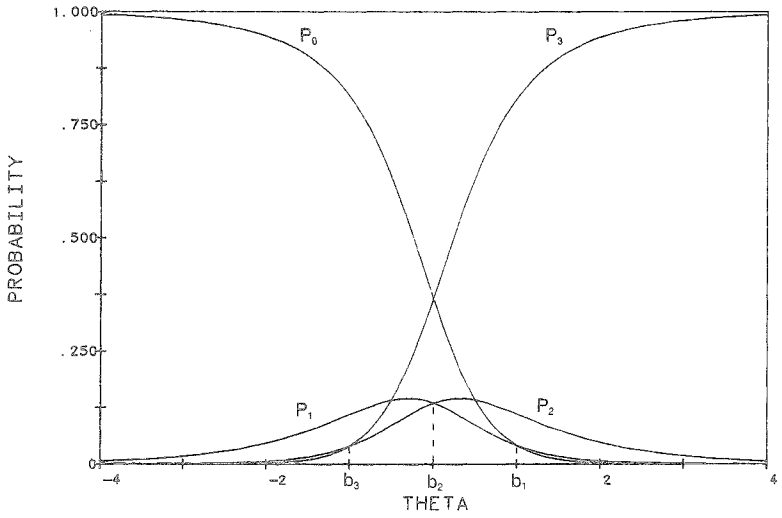
$$[(27)^{1/3} \times 4] - 3$$

is a good example of an item with a reversed ordering of the step estimates. The first step involves the cube root, which is more difficult than the second and third steps, which involve multiplication and subtraction, respectively. Similarly, the second step is more difficult than the third step. Thus, for this item, the step estimates would be ordered from most difficult to easiest. Note that this type of item, for all practical purposes, reduces the item to the two category scores of 0 and 3. This is because a person who successfully completes the first step has a high probability of successfully completing the remaining two steps. The operating characteristic for such an item with step estimates of 1.0, 0.0, and −1.0 is depicted in Figure 9. Note that while the probabilities of receiving a category score of 1 or of 2 are unlikely, they are still possible.

Items with step estimates that are in reversed sequential order may not occur very often in real life, but it seems reasonable that an item with these step characteristics could be obtained. Thus, the finding that the most peaked item information functions were obtained for items with step estimates that are ordered from most difficult to easiest may be useful in certain applied situations. For example, if a short test is needed that will discriminate among examinees around a given cutting point on the θ scale, it would be desirable to construct the items such that the steps for each item would be ordered from most difficult to easiest. In addition, each item should have its first step value slightly above the cutting point and its last step value slightly below the cutting point. Such an ordering of the steps would yield more information than any other ordering of the steps, and thus the information of the test will be highly peaked in the area of the θ scale close to the desired cutting point.

The relative efficiency analyses demonstrated the effects of various test revisions on the precision of measurement of the resulting tests relative to the

Figure 9
Operating Characteristic Curves for a Four-Category Partial Credit Item with Step Parameters in Reversed Sequential Order



base test. As expected, the effects of various test revisions on relative efficiency could be predicted from the step difficulty estimates. Items with various step estimate characteristics could be ordered in terms of the magnitude of the change in efficiency of the revised test relative to the base test for θ levels in the range of the step estimates. Items with step estimates that spanned a small range had the largest effect on the efficiency of the revised test relative to the base test. Nonsequential ordering of these step estimates had a slightly greater impact on the efficiency of the revised test than did step estimates that were in sequential order. The large-range item with sequential ordering of the step estimates had the smallest effect on the efficiency of the revised test relative to the base test. Finally, the relative efficiency comparison of the revised test that included only difficult items demonstrated the usefulness of information in the revision of a test for a given measurement situation.

Collectively, the findings of this research revealed that item information functions for the partial credit model can be useful in some applications of the model. Koch and Dodd (1985, 1986) have successfully used item information for the selection of items in simulated computerized adaptive testing sessions with real and simulated datasets that were calibrated according to the partial credit model. Target information functions could be used to develop parallel forms of a test as well as to aid in redesigning a test to yield high precision of measurement where it is most needed. The use of relative efficiency analyses could also allow for the comparison and selection of the "best" test for a given measurement situation.

The usefulness of information functions in the application of the partial credit model is not restricted to item or test selection. The present findings could prove useful in the actual construction of test items. For example, a test item could be constructed so that its latter steps are easier than its earlier steps. Such an item will yield very high precision of measurement for a given range on the θ scale.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent trait ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Dodd, B. G. (1985). Attitude scaling: A comparison of the graded response and partial credit latent trait models (Doctoral dissertation, University of Texas at Austin, 1984). *Dissertation Abstracts International, 45,* 2074A.

Dodd, B. G., & Koch, W. R. (1985, April). *Item and scale information functions for the partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7,* 15–32.

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

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Masters, G. N. (1981, April). *A Rasch model for partial credit scoring.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In C. K. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing.* Washington DC: U.S. Government Printing Office.

Thissen, D. M. (1976). Information in wrong responses to Raven progressive matrices. *Journal of Educational Measurement, 13,* 201–214.

Author's Address

Send requests for reprints or further information to Barbara G. Dodd, Measurement and Evaluation Center, The University of Texas at Austin, Box 7246, Austin TX 78713, U.S.A.