ED 386 469 TM 023 840

AUTHOR Lecointe, Darius A.
TITLE How the Collapsing of Categories Impacts the Item Information Function in Polytomous Item Response Theory.
PUB DATE Apr 95
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDkS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; Computer Simulation; Difficulty Level; Interrater Reliability; *Item Response Theory; *Responses; Testing; *Test Items
IDENTIFIERS *Information Function (Tests); Partial Credit Model; *Performance Based Evaluation; Polytomous Items

ABSTRACT

The purpose of this Item Response Theory study was to investigate how the expected reduction in item information, due to the collapsing of response categories in performance assessment data, was affected by varying testing conditions: item difficulty, item discrimination, inter-rater reliability, and direction of collapsing. The investigation used copulas in computer simulations of performance assessment batteries with varying combinations of item characteristics within Muraki's Generalized Partial Credit Model. Only two of the significant contrasts that were detected were of practical importance. The results appear to indicate that the expected reduction in information due to the collapsing of categories is not affected by any of the testing conditions simulated in this study. Consequently, a practitioner may combine low-frequency categories with adjacent categories without any significant adverse effect on the information provided by the test items. (Contains 1 figure, 9 tables, and 19 references.) (Author)

How the Collapsing of Categories Impacts the Item Information Function in Polytomous Item Response Theory

Darius A. Lecointe
Educational Testing Service, Princeton, New Jersey

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1995

Abstract

The purpose of this Item Response Theory study was to investigate how the expected reduction in item information, due to the collapsing of response categories in performance assessment data, was affected by varying testing conditions: item difficulty, item discrimination, inter-rater reliability, and direction of collapsing. The investigation used copulas in computer simulations of performance assessment batteries with varying combinations of item characteristics within Muraki's Generalized Partial Credit Model. Only two of the significant contrasts that were detected were of practical importance. The results appear to indicate that the expected reduction in information due to the collapsing of categories is not affected by any of the testing conditions simulated in this study. Consequently, a practitioner may combine low-frequency-categories with adjacent categories without any significant adverse effect on the information provided by the test items.

How the Collapsing of Categories Impacts the Item Information Function in Polytomous Item Response Theory

Introduction

Performance assessments have a history dating back as far as the ancient Greeks, and they have been receiving increasing attention over the past fifty years (Davey, 1991). Much of this attention is because performance assessments facilitate the assessment of skills that cannot be adequately evaluated with paper-and-pencil assessment strategies (Oosterhof, 1990; Stiggins, 1987). Recognizing the growing importance of performance assessments, the National Council on Measurement in Education (NCME) published an instructional module to assist educators and assessment specialists in designing and developing performance assessments (Stiggins, 1987).

Polytomous Item Response Theory models are also receiving increasing attention, emerging as the model of choice for the analysis of the type of data obtained from performance assessments. A number of studies (De Ayala, Dodd, & Koch, 1989; Ferrara & Walker-Bartnick, 1989; Muraki & Wang, 1992; and Phillips, Mead, & Ryan, 1983) have focussed on assessments of writing achievement. These studies have demonstrated that Master's Partial Credit Model may be used in the analysis of performance assessment data as well as of cognitive data (Brown, 1992; Huynh & Ferrara, 1992).

When polytomous scoring is used one must consider the incidence of low frequency categories (LFC's) occurring naturally in the data or as a result of combining the assessments of multiple raters. When low frequency categories are present a practitioner has the option of using the data as they have been gathered or of collapsing the LFC's into adjacent categories (Brown, 1991).

Dodd and Koch (1986), Muraki (1992), and Allen (1992) demonstrated that a reduction in the number of categories, either through the use of items with fewer categories or through the collapsing of categories, does affect the information function. Generally, the amount of information provided increases as the number of categories increases. However, Allen found that the item information function curve had a higher peak (maximum information) after collapsing, and the amount of information (total information) increased over the θ scale. She attributed this discrepancy to the effect of collapsing categories at the lower end of the scale.

This study attempts to clarify this observed discrepancy, and is another venture in the continuing investigation of the relationship between the collapsing of categories and the information function.

Method

Data Generator

A Fortran program was written to simulate the combined scores of two raters. The model used was the rating scale version of Muraki's Generalized Partial Credit Model. The Generalized Partial Credit Model specifies that, for an item with m categories, an examinee's probability of success on category k of an item, given successful completion of the previous (easier) categories, is given by the following equation:

$$\pi_{nik} = \frac{\exp\sum_{j=0}^{k} Da_i[\theta_n-(b_i+\tau_j)]}{\sum_{h=0}^{m}\exp\sum_{j=0}^{h} Da_i[\theta_n-(b_i+\tau_j)]} \qquad k=0,1,\ldots m_i \ , \qquad (1)$$

where a_i is the slope parameter, b_i is the item location parameter, and τ_j is the category or threshold parameter which, in the rating scale model, is the same for all items. D is a scaling constant that determines the metric of the θ scale. D=1.7 puts the θ scale on the same metric as the normal ogive model, D=1.0, the value used in this study, expresses the model in the logistic metric (Muraki & Bock, 1992b).

The program simulated a twelve-item test and generated the scores for one thousand examinees. All the items in each simulated test had the same level of item discrimination and three different levels of difficulty were simulated within each test (four items at each level of difficulty). Step-difficulty parameters for four-step items were obtained by setting three threshold values at -1.0, 0.0, and 1.0.

Examinee ability values were randomly selected from the standard normal distribution and, using pre-specified item characteristic parameters, category probability values were calculated. The procedure described by Walker-Bartnick (1990) for transforming these probabilities into response data was modified to provide the correlated bivariate data. Under the basic procedure for simulating polytomous data, cumulative probability intervals (C_1, C_2, . . . C_k, for an item with scores ranging from 0 to k) are compared with a random variate (r) drawn from a uniform distribution on the [0,1] interval to simulate the score on one item for each examinee, so that

$$x = k: C_{j-1} < r \leq C_j \qquad (j=1,2,\ldots k) \qquad (2)$$
$$x = 0: otherwise$$

In order to pre-specify the inter-rater reliability of the simulated raters' scores, two correlated uniform distributions were used in the modification. This combination of two correlated uniform distributions in the form of a bivariate distribution with marginals that are uniform on the [0,1] interval, is called a copula. For each item, a copula with a size equal to the number of examinees to be simulated was first generated. Walker-Bartnick's procedure was then applied to each pair of random variates from the copula, resulting in an array of scores that simulated the ratings, for one item, granted by two raters to each examinee. During each run the computer program replicated this procedure twelve times, to simulate a twelve-item test.

The kappa index (Fleiss, 1981) was calculated to determine the simulated inter-rater reliability of the generated scores. The correlation level of each copula was manipulated to produce scores with the required level of kappa.

Research and Test Design

The research design used in this study was a 3 x 3 x 2 x 2 repeated measures design. Four experimental factors were selected. They were: 1) item difficulty at three levels (-1.0, 0.0, 1.0), 2) item discrimination at three levels (0.4, 0.9, 1.6), 3) inter-rater reliability at two levels (Low, High),[1] and 4) direction of collapsing at two levels (upward and downward).[2]

Two of the variables, inter-rater reliability and item discrimination, were held constant in each test. This design arrangement required only six unique test types to encompass the four experimental factors. The twelve items in each test were grouped into three blocks, each at one of the three experimental levels of item difficulty.

[1] The mean kappa values were .74 (standard deviation: .04, range: from .63 to .82) and .88 (standard deviation: .02, range: from .83 to .94) for the "low" and "high" levels, respectively. The "low" level more accurately reflects the values generally found in the literature.

[2] The low-frequency-categories were collapsed with categories above and below them for the "direction of collapsing" variable.

The dependent variable was the information provided by each item, and the sampling unit was the test item. A preliminary power analysis indicated a total of sixteen items in each cell of the research design. Since each difficulty level in a test was assigned to four test items, each test simulation was replicated four times to produce a total of 24 tests, with 288 items in the data set.

PARSCALE (Muraki & Bock, 1992a) was used for the test calibrations, and the information files from PARSCALE were used to provide the dependent variable values. After the original data were calibrated the low frequency categories in each of the 288 items were identified. The tests were then recalibrated twice: 1) after the low frequency categories had been collapsed with adjacent categories above them (upward), and 2) after the low frequency categories had been collapsed with adjacent categories below them (downward). Because fewer categories are expected to provide less information, two dependent variables, the difference between the "original" information and each of the "collapsed" information data, were computed. These two variables represented the repeated measures component of the design.

Preliminary Analysis and Data Inspection

Preliminary analyses were conducted to (1) confirm a common scale for the information data, (2) identify outliers in the data, and (3) ensure that the data conformed to theoretical expectations.

Common Information Scale

For some of the generated tests only a subset of the test items met the collapsing criterion. Because this study was, in part, based on the assumption that the collapsing of categories does not change the underlying scale of the information data, the estimated a and b parameters, from the second and third calibrations, of items that were not collapsed were compared with the estimated parameters from the first calibration of the original items. No statistical difference was detected between the estimated parameters from the first calibration and those from the second and third. The assumption that the information values from the collapsed and uncollapsed items were on a common scale was upheld, indicating that the statistical analysis of the information from collapsed and non-collapsed items was justified.

Outliers in the Data

Seven items, from five different tests, whose information values before or after collapsing did not meet expectations from information theory were flagged as outliers. Data from the first calibration showed that four items of "medium" item discrimination provided more maximum information than items with "high" item discrimination, and a fifth item with "high" inter-rater reliability had more than three times the maximum information from any of the other items with similar discrimination and difficulty values. After collapsing, one item had zero maximum information and a second had more than six times the maximum information from any of the other items with the same inter-rater reliability value. The same trends were observed with the total information data, although the factors were smaller.

The test calibration results for these flagged items were evaluated further. Although all the tests converged through the EM-cycle phase, they also displayed some divergence during the Newton-cycle phase. (This is the second and last iteration phase in PARSCALE). This was not considered to be sufficient cause to drop the data from those tests because other tests that did not have items with irregular information function curves also displayed some divergence during that phase. Two other phenomena were noted: 1) all of the calibration runs terminated without any error messages, but although the program was allowed to default to a convergence criterion of .001, a number of tests had a final change that was greater than .001; 2) only two of the flagged items had final parameter estimates which were much different from the estimates given after the EM-cycle phase.

The effect of manually setting the convergence criterion at the default value specified in the program manual was investigated by recalibrating all the tests that had flagged items, along with a few of the other tests. The following results were observed:

1) the largest change always satisfied the convergence criterion,
2) the divergence problem did not always disappear,
3) the irregularities in the information curves from the flagged items disappeared,
4) the new calibrations resulted in only minimal changes in information for tests without flagged items,
5) other items in the same tests as the flagged items also showed only minimal changes in their information functions.

On the basis of these observations only the seven flagged items were dropped from the data set. This represented a 2.4% loss of data. The remaining 281 items were analyzed with a 3 x 3 x 2 x 2 repeated measures analysis.

Conformity to Theoretical Expectations

The information data from the original data calibrations are displayed in Tables 1 and 2. The tables show that the generated data conformed to the requirements of information theory. First, information is mathematically represented as a function of the square of the discrimination parameter. As expected, both the maximum and total information values increased as item discrimination increased.

****

Insert Tables 1 and 2 about here

****

Secondly, with other factors kept constant, information is expected to be the same for the two symmetric item difficulty values, -1.0 and 1.0, specified in this study. Within the bounds of sampling error those expectations were met. The final observation from the baseline data was the relationship between inter-rater reliability and information. Although performance assessment theory indicates that higher levels of inter-rater reliability are more desirable, the data show that there was no marked numerical difference between the information provided at the two simulated levels of inter-rater reliability.

Statistical Criteria

The significance level for all tests, effects and contrasts, was set at $\alpha=.05$, but there was not a strict adherence to the rule. The practical importance of the significant and notable effects was judged in terms of the percentage decrease from baseline for the effects, or the difference in percentages for the contrasts. A decrease or difference of 10% was the level at which an effect or contrast was considered to be of practical importance.

Results

Overall Effects

The statistical package, SPSS/PC+, was used to conduct the repeated measures analyses. In SPSS/PC+, when the original variables in a repeated measures analysis are difference scores, as in this study, the test of the constant corresponds to the test that "there has been no overall change from baseline" (Norusis, 1990, p. B-116). The baseline was the mean maximum and total item information provided by the unmodified test items (see Tables 1 and 2). The test of the constant was non-significant ($F_{(1,263)}=.66$, $p=.419$) for the maximum information data, and significant ($F_{(1,263)}=34.98$, $p<.001$) for total information data. Although the collapsing of categories did not result in a significant overall decrease in maximum information, there was a significant overall reduction in total information. These results indicate that, whereas the peak of the information curve may not change appreciably, the tails of the curve may be expected to be thicker after categories have been collapsed.

Figure 1 contains plots of the item information curves for one of the items in the data set, before and after categories were collapsed; relevant details are summarized in Table 3.

****
Insert Figure 1 and Table 3 about here
****

First, the information curve registered a shift (to the left) only after categories were collapsed upward; this shift corresponds to a reduction in item difficulty after categories were collapsed upward. Second, the information curve was flatter than the original curve after categories were collapsed upward, and more peaked after categories were collapsed downward; this corresponds to a decrease in maximum information after upward collapsing, and an increase after downward collapsing of categories. Third, the area under the information curve was lesser for the upward collapsed data (8.94), and greater for the downward collapsed data (9.77) when compared with the original data (9.36). For this item it can be inferred that more information was provided, if categories were collapsed upward, for examinees at the lower end of the ability scale. At the extremes of the scale the information from categories collapsed downwards was the same as for the original items. However, across the mid-range of the ability scale downward collapsing produced more information than the original items.

Table 3 shows that although the simulated difficulty value was 0.00 the parameter estimates were slightly greater. This is because the item characteristics from the single rater scores were changed when the two raters' scores were combined into one score.

Overall, the combining of the raters' scores caused only a minimal change from the simulated difficulty valuc, but there was marked change in the range of step difficulty values. The range showed a marked increase after the ratings were combined, and steps that originally were ordered in difficulty were replaced by reversals--alternating between very easy steps and very difficult steps. This pattern of reversals produced a saw-toothed pattern in the frequency distributions of the response categories that, in this study, was desired to ensure that the low frequency categories were distributed along the range of the rating scale.

The effective discrimination of the items resulting from the combination of the two simulated raters' scores was also less than the values assigned to the original items. On average, the input item discrimination parameters of .4, .9 and 1.6 were estimated as .1, .3, and .5, respectively, reflecting cuts of about one third. It also was observed that the degree of association, inter-rater reliability, between the two sets of scores that were combined did not appear to affect the estimated discrimination of the items.

Within-Subjects Effects

None of the effects involving the direction of collapsing was significant. The change in maximum or total information was the same after categories were collapsed upward or downward. This result may be explained by the fact that in this study the final number of categories remained the same under both collapsing conditions. Circumstances under which the two directions of collapsing result in different numbers of categories may produce different results.

Between-Subjects Effects

Both the maximum and total information analyses had two between-subjects interaction effects that were significant: inter-rater reliability x item discrimination interaction ($F_{(2,263)}=4.31$, $p=.014$ [maximum] and $F_{(2,263)}=2.88$, $p=.058$ [total]) and item discrimination x item difficulty interaction ($F_{(2,263)}=5.41$, $p<.001$ [maximum] and $F_{(2,263)}=5.41$, $p=.006$

[total]). The effects were studied by focussing on the contrasts from the simple, simple main effects.

Post Hoc Tests and Practical Importance

In the absence of a significant "direction of collapsing" effect on either maximum information or total information, the contrasts were computed from the average of the information data from upward and downward collapsing. The relevant averages and contrasts are presented for maximum information in Tables 4, 5, and 6, and, for total information, in Tables 7, 8, and 9. The reported percentages represent the percentage change from the baseline values in Table 1 (maximum information) and Table 2 (total information).

****
Insert Tables 4, 5, 6, 7, 8, and 9 about here
****

The confidence intervals displayed in Tables 5 and 6, and Tables 8 and 9 are not based on data representing actual mean maximum or total information, but the mean decrease in maximum or total information. These intervals demonstrate whether the decrease in maximum or total information was different at various levels of the relevant factor. Because an increase in information is a desirable outcome contrasts involving increases in information are only noted, but are not discussed or evaluated.

****
Insert Tables 10 and 11 about here
****

The data on the contrasts clearly show that even the significant contrasts are not large. The picture is clearer when displayed in terms of percentages (Tables 10 and 11) In only a few situations were significant contrasts observed. All but one of these situations involved the "high" level of the relevant factor or testing condition. The exception was in the total information data where the significant contrasts between "high" level and both the "low" and "medium" levels of item discrimination were observed under the condition of "low" inter-rater reliability.

Only in the maximum information data did the collapsing of categories result in an increase in information. The increases ranged from .1% over the baseline value (Table 4) to a 10.2% increase over the baseline value for the effects associated with the significant contrasts. The data in Tables 10 and 11 also show that, on average, the effects and contrasts were greater for the total information data than for the maximum information data.

Finally, only two contrasts of practical importance were observed, and both were in the total information data: under conditions of "high" item discrimination the collapsing of categories resulted in an impact on total information that was 12.1 percentage points greater at the "high" level of item difficulty than at the "medium" level, and under conditions of "high" item difficulty the impact was 10.5 percentage points greater at the "high" level of item discrimination than at the "medium" level.

Discussion

Aside from the major issues addressed by the results several side issues that were not part of the primary statistical treatment emerged, albeit within the context of the conditions simulated in this study. First, it was demonstrated that higher levels of inter-rater reliability may not necessarily result in higher values of item information. While the case can be made that specialized training of raters is recommended to produce higher levels of inter-rater reliability that would overcome the inherent subjective nature of these judgement-based scores, these results appear to indicate that there is not much to be gained in trying to improve inter-rater reliability above the levels presently reported in the literature, represented by the "low" level of the IRR factor.

Second, the degree of association, inter-rater reliability, between the two sets of scores that were combined did not appear to affect the estimated discrimination of the items. In light of the first, this finding is significant. There does not appear to be any justification for expenditure of resources to increase present levels of inter-rater reliability if that expenditure cannot be expected to produce more information or result in better discrimination between high and low scoring candidates.

The third finding is that the merger of two raters' scores into one score reduces the item characteristics of the original items. The diminution effect of the combination of two raters' scores is more pronounced on item discrimination than on item difficulty. But the effect indicates that developers of assessment instruments requiring subjective assessments by raters should be aware that the original item characteristics may need to be set at much higher values than their target values for those characteristics.

Even though these results are more directly relevant to the conditions simulated here, the incidence of low frequency categories is not restricted solely to those conditions where the ratings of two or more raters with a high level of inter-rater reliability are combined into one score. For example, they may occur in a situation where the work of the majority of examinees is of such quality that certain points on the scale are barely used by the raters. The obvious question is whether collapsing may have different effects depending on the circumstances that made the collapsing necessary. There does not appear to be any reason why this should be the case. This concern may need further research, but this researcher believes that the results of this study may be applied to other conditions where LFC's are present.

Not only were there only two contrasts large enough to be considered of practical importance, but the overall impact of collapsing, as measured by the test of the constant, was not significant in the maximum information data, and the impact was not of practical importance in the total information data.

The observed significant effects must be interpreted in light of the fact that tests used in practical situations are usually heterogenous, composed of items with varying combinations of item parameters. It is extremely unlikely that a practitioner would have to contend with a test in which the testing condition is overwhelmingly similar to those simulated conditions associated with the two contrasts of practical importance.

Consequently, the decision on how to deal with LFC's in polytomous test data is not a high stakes one. The practitioner may collapse the LFC's with adjacent categories knowing that this decision will not produce adverse effects on the information function. By the same token the practitioner who decides to leave the data as they are may do so.

Another consideration concerns the fact that the two contrasts of practical importance were observed in the total information data and not in the maximum information data. Of course, maximum item information is more directly relevant to the estimated person parameters because it is related to the precision of the person parameter estimates. This only further strengthens the conclusions that have been drawn.

Clearly, these results have provided support for Allen's observation that the reduction in the number of categories may, albeit unexpectedly, increase the information provided by an item. They also provide some direction for practitioners who may be concerned about the impact on item information of a decision to collapse LFC's that occur in their data.

References

Allen, N. L. (1992, February). An application of BILOG/PARSCALE to the 1990 science cross-sectional items. A paper presented at Design and Analysis Committee of the National Assessment of Educational Progress, Alexandria, VA.

Brown, R. L. (1991). The effect of collapsing ordered polytomous scales on parameter estimates in structural equation measurement models. Educational and Psychological Measurement, 51(2), 317-328.

Brown, W. L. (1992, April). Analysis of Spring 1990 field tryouts of MEAP essential skills mathematics test items using a Partial Credit model. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Davey, B. (1991). Evaluating teacher competence through the use of performance assessment tasks: An overview. Journal of Personnel Evaluation in Education, 5(2), 121-132.

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1989, March). A comparison of the graded response and the partial credit models for assessing writing ability. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Dodd, B. G., & Koch, W. R. (1986, April). Relative efficiency analyses for the partial credit model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Ferrara, S., & Walker-Bartnick, L. (1989, March). Constructing an essay prompt bank using the partial credit model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. New York: John Wiley & Sons.

Huynh, H. & Ferrara, S. (1992, April). A comparison of equal percentile and partial credit equatings for open-ended examinations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Unpublished manuscript, Princeton, NJ: ETS.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159-76.

Muraki, E., & Bock, R. D. (1992a). Parscale: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks [Computer Program]. Chicago: Scientific Software, Inc.

Muraki, E., & Bock, R. D. (1992b). Parscale: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks [Program manual]. Chicago: Scientific Software, Inc.

Muraki, E., & Wang, M. (1992, April). Issues relating to the marginal maximum likelihood estimation of the Partial Credit model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Norusis, M. J. (1990). SPSS/PC+ Advanced Statistics 4.0. Chicago: SPSS Inc.

Oosterhof, A. C. (1990). Classroom applications of educational measurement. Columbus, Ohio: Merrill Publishing Company.

Phillips, G., Mead, R. & Ryan, J. (1983). Use of the general polychotomous form of the Rasch model as a means of calibrating and equating indirect and direct assessments of writing achievement. Paper presented at the National Council on Measurement in Education, Montreal.

Stiggins, R. J. (1987). NCME instructional module on design and development of performance assessments. Educational Measurement Issues and Practice, 6(3), 33-42.

Walker-Bartnick, L. A. (1990). An investigation of factors affecting invariance of item parameter estimates for the partial credit model. Doctoral Dissertation, University of Maryland.

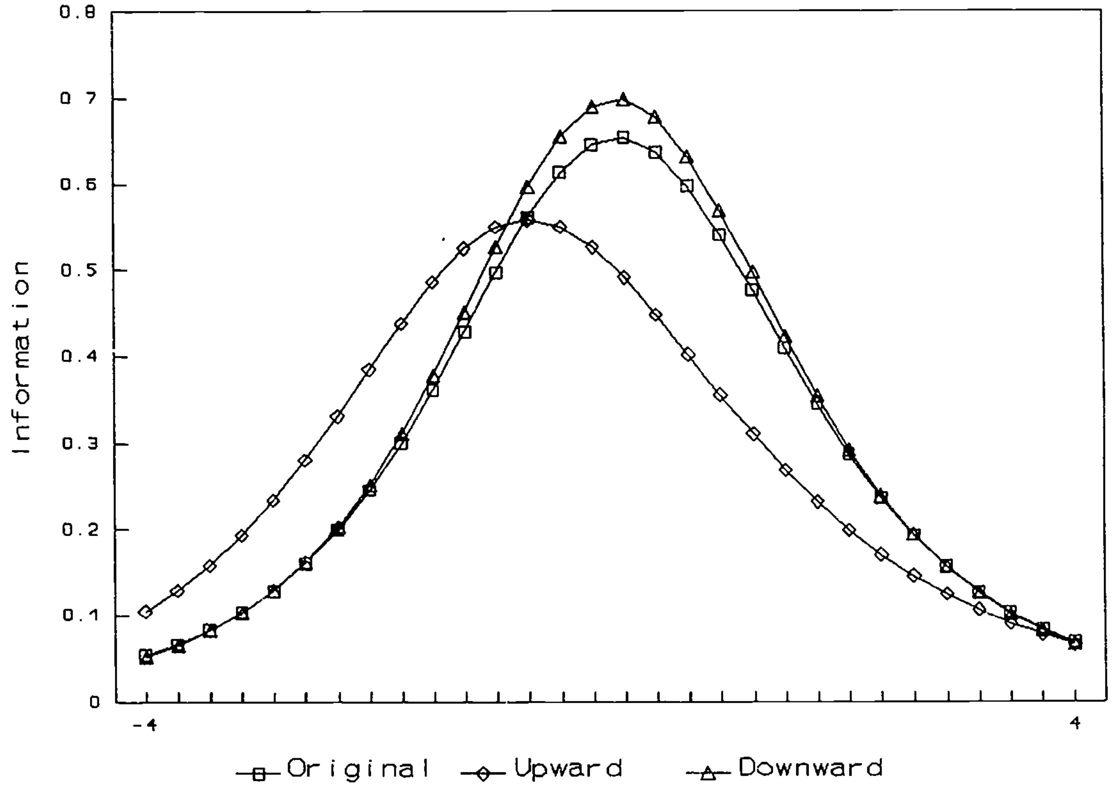

Figure 1 Item Information Function: Comparison of Original and Collapsed Categories (a=.9, b=0.0,IRR=Low)

Table 1

Mean Maximum Information From Unmodified Responses+ (n=16)

IRR*	Item Discrimination	Item Difficulty			Mean
		low	medium	high	
low	low	.19 (.05)	.18@ (.04)	.18 (.06)	.18 (.05)
	medium	.74# (.39)	.74% (.43)	.67 (.25)	.72 (.35)
	high	1.62 (.16)	1.68 (.18)	1.77 (.18)	1.69 (.18)
	Mean	.86 (.66)	.89 (.69)	.87 (.69)	.87 (.68)
high	low	.17 (.04)	.17 (.04)	.16 (.04)	.17 (.04)
	medium	.68 (.10)	.73 (.08)	.78 (.08)	.73 (.09)
	high	1.68 (.50)	1.57 (.36)	1.89 (.52)	1.72 (.48)
	Mean	.84 (.70)	.81 (.61)	.94 (.78)	.87 (.70)
Mean		.85 (.67)	.85 (.65)	.91 (.74)	.87 (.69)

+ Values in parentheses are the standard deviations

* IRR: Inter-Rater Reliability

@ n=15

n=13

% n=14

Table 2

Mean Total Information From Unmodified Responses[+] (n=16)

IRR[•]	Item Discrimination	Item Difficulty			Mean
		low	medium	high	
low	low	3.62 (.81)	3.65[@] (.59)	3.53 (.87)	3.60 (.75)
	medium	9.53[#] (4.20)	9.14[%] (3.91)	9.03 (3.12)	9.22 (3.65)
	high	18.59 (1.34)	19.10 (1.44)	19.75 (1.34)	19.14 (1.43)
	Mean	10.65 (6.85)	10.85 (6.98)	10.77 (7.09)	10.76 (6.92)
high	low	3.35 (.62)	3.51 (.55)	3.26 (.54)	3.37 (.57)
	medium	9.30 (.87)	9.95 (.69)	10.23 (.69)	9.83 (.84)
	high	18.36 (4.36)	17.91[@] (3.30)	20.62 (4.05)	18.99 (4.04)
	Mean	10.34 (6.73)	10.30 (6.21)	11.37 (7.58)	10.67 (6.84)
Mean		10.49 (6.75)	10.57 (6.57)	11.07 (7.30)	10.71 (6.87)

[+] Values in parentheses are the standard deviations

[•] IRR: Inter-Rater Reliability

[@] n=15

[#] n=13

[%] n=14

Table 3

Item Characteristics Before and After Category Collapsing

Characteristic	Original Categories	Upward Collapsing	Downward Collapsing
Difficulty (0.00)	.15	.13	.15
Maximum information	.65	.56	.70
Total Information	9.36	8.94	9.77

Table 4

Two-Way Tables of Means For Maximum Information:

Average Decrease From Upward and Downward Collapsing

Item Difficulty	Item Discrimination			Mean
	low	medium	high	
low	.01 (32)[@] [4.5][#]	-.00 (29) [-.6]	.04 (32) [2.5]	.02 (93) [1.9]
medium	.00 (31) [1.7]	-.00 (30) [-.1]	-.17 (31) [-10.2]	-.06 (92) [-6.5]
high	.01 (32) [6.4]	.02 (32) [2.9]	.19 (32) [10.2]	.07 (96) [8.0]
IRR				
low	.01 (47) [4.9]	-.02 (43) [-2.4]	-.07 (48) [-4.2]	-.03 (138) [-3.1]
high	.01 (48) [3.6]	.03 (48) [3.6]	.12 (47) [6.9]	.05 (143) [5.8]
	.01 (95) [4.0]	.01 (91) [.8]	.02 (95) [1.4]	.01 (281) [1.4]

[+] Negative values represent an increase from baseline values

[@] Cell size

[#] Percentage decrease from baseline

Table 5

95% Confidence Intervals on Item Difficulty and Inter-Rater Reliability

Contrasts: Maximum Information

Contrast	Factor	Point Estimate	Std. Error	Confidence Interval
Diff	Disc			
low-medium	low	.01	.084	-.19, .21
	medium	-.00	.086	-.20, .20
	high	.21*	.084	.01, .41+
low-high	low	.00	.083	-.19, .19
	medium	-.02	.085	-.22, .18
	high	-.15*	.083	-.34, .04
medium-high	low	-.01	.084	-.21, .19
	medium	-.02	.084	-.22, .18
	high	-.36*	.084	-.56, -.16
IRR				
low-high	low	.00	.068	-.13, .13
	medium	-.05	.070	-.19, .09
	high	-.19*	.068	-.32, -.06

C_v=2.34 (Diff); 1.96 (IRR)

* point estimate considered not equal to zero

+ significant contrasts indicated by Bonferroni tests

Table 6

95% Confidence Intervals on Item Discrimination Contrasts: Maximum Information

Contrast	Factor	Point Estimate	Std. Error	Confidence Interval
Disc	Diff			
low-medium	low	.01	.085	-.19, .21
	medium	.00	.085	-.20, .20
	high	-.01	.083	-.20, .18$^+$
low-high	low	-.03	.083	-.22, .16
	medium	.17*	.084	.03, .37
	high	-.18*	.083	-.37, -.01
medium-high	low	-.04	.085	-.24, .16
	medium	.17*	.085	.03, .37
	high	-.17*	.083	-.36, -.02
Disc	IRR			
low-medium	low	.03	.070	-.13, .19
	high	-.02	.068	-.18, .14
low-high	low	.08*	.068	-.08, .24
	high	-.11*	.068	-.27, .05
medium-high	low	.05*	.070	-.11, .21
	high	-.09*	.068	-.25, .07

C_v=2.34

* point estimate considered not equal to zero

$^+$ significant contrasts indicated by Bonferroni tests

Table 7

Two-Way Tables of Means For Total Information:

Average Decrease From Upward and Downward Collapsing

	Item Discrimination			
Item Difficulty	low	medium	high	Mean
low	.15 (32)@ [4.3]#	.35 (29) [3.8]	1.51 (32) [8.2]	.68 (93) [6.5]
medium	.36 (31) [1.6]	.21 (30) [2.2]	.46 (31) [2.5]	.24 (92) [2.3]
high	.21 (32) [6.0]	.40 (32) [4.1]	2.95 (32) [14.6]	1.18 (96) [10.7]
IRR				
low	.18 (47) [5.1]	.37 (43) [4.1]	2.28 (48) [11.9]	.97 (138) [9.0]
high	.09 (48) [2.8]	.28 (48) [2.8]	1.01 (47) [5.3]	.46 (143) [4.3]
	.14 (95) [4.0]	.32 (91) [3.4]	1.65 (95) [8.6]	.71 (281) [6.6]

@ Cell size

# Percentage decrease from baseline

Table 8

95% Confidence Intervals on Item Difficulty and Inter-Rater Reliability

Contrasts: Total Information

Contrast	Factor	Point Estimate	Std. Error	Confidence Interval
Diff	Disc			
low-medium	low	.09	.704	-1.56, 1.74
	medium	.14	.728	-1.56, 1.84
	high	1.05	.704	-.60, 2.70[+]
low-high	low	-.06	.699	-1.70, 1.58
	medium	-.05	.717	-1.73, 1.63
	high	-1.44	.699	-3.08, .20
medium-high	low	-.15	.704	-1.80, 1.50
	medium	-.19	.710	-1.85, 1.47
	high	-2.49*	.704	-4.14, -.84
IRR				
low-high	low	.09	.573	-1.03, 1.21
	medium	.09	.587	-1.06, 1.24
	high	1.27*	.573	.15, 2.39

C_v=2.34 (Diff); 1.96 (IRR)

* Point estimate considered not equal to zero

[+] Significant contrast indicated by Bonferroni tests

Table 9

95% Confidence Intervals on Item Discrimination Contrasts: Total Information

Contrast	Factor	Point Estimate	Std. Error	Confidence Interval
Disc	Diff			
low-medium	low	-.20	.717	-1.88, 1.48$^+$
	medium	-.15	.716	-1.83, 1.53
	high	-.19	.699	-1.83, 1.45
low-high	low	-1.36	.699	-3.00, .28
	medium	-.40	.710	-2.06, 1.26
	high	-2.74*	.699	-4.38, -1.10
medium-high	low	-1.16	.717	-2.84, .52
	medium	-.25	.716	-1.93, 1.43
	high	-2.55*	.699	-4.19, -.91
Disc	IRR			
low-medium	low	-.19	.590	-1.57, 1.19
	high	-.19	.570	-1.52, 1.14
low-high	low	-2.10*	.573	-3.36, -.84
	high	-.92	.573	-2.18, .34
medium-high	low	-1.91*	.587	-3.28, -.54
	high	-.73	.573	-2.07, .61

C_v=2.34

* Point estimate considered not equal to zero

$^+$ Significant contrast indicated by Bonferroni tests

Table 10

Contrasts and Effects for Maximum Information in Percentages

Testing Condition		Contrast			Effect
		Low	Medium	High	Decrease (Increase)
			Item Difficulty		
High Discrimination	Low	--	na[a]	7.7	2.5
	Medium		--	na	(10.2)
	High			--	10.2
			Inter-rater Reliability		
High Discrimination	Low	--	▨	na	(4.2)
	High		▨	--	6.9
			Item Discrimination		
Medium Difficulty	Low	--	na	na	1.7
	Medium		--	na	(.1)
	High			--	(10.2)
High Difficulty	Low	--	ns[b]	3.8	6.4
	Medium		--	7.3	2.9
	High			--	10.2
High IRR	Low	--	ns	3.3	3.6
	Medium		--	3.3	3.6
	High			--	6.9

[a] A contrast involving an increase in information

[b] Not significant at the .05 level of significance

Table 11

Contrasts and Effects for Total Information in Percentages

Testing Condition		Contrast			Effect
		Low	Medium	High	Decrease
Item Difficulty					
High Discrimination	Low	--	ns[a]	ns	8.2
	Medium		--	12.1	2.5
	High			--	14.6
Inter-rater Reliability					
High Discrimination	Low	--		6.6	11.9
	High			--	5.3
Item Discrimination					
High Difficulty	Low	--	ns	8.6	6.0
	Medium		--	10.5	4.1
	High			--	14.6
Low IRR	Low	--	ns	6.8	5.1
	Medium		--	7.8	4.1
	High			--	11.9

[a] Not significant at the .05 level of significance