DOCUMENT RESUME

ED 365 705 TM 020 891

AUTHOR Oshima, T. C.; And Others
TITLE Differential Item Functioning with a Criterion-Referenced Test: Use of Limited Closed-Interval Measures.
PUB DATE Apr 93
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Criterion Referenced Tests; *Cutting Scores; Data Analysis; Evaluation Methods; *Item Bias; Item Response Theory; *Test Construction
IDENTIFIERS *Limited Closed Interval Measures

ABSTRACT

The purpose of this study was to introduce a procedure to detect differential item functioning (DIF) particularly suitable for criterion-referenced tests and to demonstrate how this approach would affect the identification of DIF items using real data sets. The procedure based on item response theory (IRT) assesses DIF at a limited closed interval of thetas at which a cutoff score fails. Illustrative data showed that identification of DIF could be quite different with this unconventional procedure as opposed to traditional DIF measures with which DIF was assessed over the entire range of ability. It was recommended that test development practitioners be actively involved in the DIF analysis to investigate not only if an item is biased, but where it may be biased. Three figures, and five tables are included. (Contains 8 references.) (Author)

Differential Item Functioning with a Criterion-Referenced Test: Use of Limited Closed-Interval Measures

T. C. Oshima

Dixie McGinty

and

Claudia P. Flowers

Georgia State University

Paper presented at the annual meeting of National Council on Measurement in Education, Atlanta, April, 1993.

Running head: DIF with a CRT test

Abstract

The purpose of this study was to introduce a procedure to detect differential item functioning (DIF) particularly suitable for criterion-referenced tests and to demonstrate how this approach would affect the identification of DIF items using real data sets. The procedure based on item response theory (IRT) assesses DIF at a limited closed interval of θs at which a cutoff score falls. Illustrative data showed that identification of DIF could be quite different with this unconventional procedure as opposed to traditional DIF measures with which DIF was assessed over the entire range of ability. It was recommended that test development practitioners be actively involved in the DIF analysis to investigate not only if an item is biased, but where it may be biased.

Differential Item Functioning with a Criterion-Referenced Test: Use of Limited Closed-Interval Measures

In recent years, the use of statistical procedures to detect differential item functioning (DIF) across two or more subgroups of examinees has become increasingly widespread. A number of different approaches to detecting DIF are currently in use, including several that are based on item response theory (IRT) (see Hills, 1989, for a review of various item bias indexes). These approaches, either IRT based or non-IRT based, generally assess bias across the entire ability range of examinees.

For a norm-referenced test, it could be argued that DIF is a concern no matter where in the ability range it occurs, since norm-referenced tests aim to measure differences among individuals at all points along the ability continuum. Developers of criterion-referenced tests, on the other hand, should be extremely concerned about DIF that occurs near the cutoff score, but much less concerned when it occurs considerably above or below the cutoff score. Though criterion-referenced tests have generally been analyzed for DIF using the same procedures that are used for norm-referenced tests, the difference in purpose of these two types of testing suggests that the issue of potential item bias on criterion-referenced tests should be addressed in a different way. While any existing bias detection method may be appropriate for use with a norm-referenced test, criterion-referenced testing calls for special attention to potential bias in the region near the cutoff

score.

With the recent development of closed-interval formulas by Kim & Cohen (1991) for measuring areas between two item characteristic curves (ICCs), the task of assessing bias over a certain ability range is as effortless as assessing bias over the entire ability range. It can be done using the user-friendly computer program IRTDIF, also developed by Kim and Cohen (1992). The purpose of their study, however, was not to introduce a small interval range as we will propose in this paper, but to introduce an alternative approach to Raju's (1988) exact area measure between two ICCs. In their study, Kim & Cohen showed that exact and closed-interval methods yielded very similar results when the interval used was $(\theta,\theta)=(-4,4)$. Though their formulas can, of course, be applied to any closed interval of interest, there has as yet been no research to determine how DIF indexes are affected when narrower closed intervals are used.

The purpose of the current study was to introduce an unconventional way of assessing item bias suitable for criterion-referenced tests and to demonstrate how this change would affect the identification of DIF items using real data sets. We proposed, for our particular test, to limit the range of ability for assessing bias to $(\theta_1, \theta_2) = (-2, 0)$, the region which contained the cut-off score. For clarity of presentation, we shall call this measure the "Q2" measure, as opposed to the conventional "Q0" measure which calculates the area between ICCs across the larger ability range (i.e., $(\theta_1, \theta_2) = (-4, 4)$). See Figure 1 for a

graphic presentation of Q0 and Q2 measures.

Insert Figure 1 about here

The current study addresses three research questions:

1. How are items classified differently with respect to DIF when the Q2 measure is applied instead of a conventional Q0 measure?

2. How does the Q2 measure correlate with other conventional DIF measures such as Raju's exact area measure and Lord's chi-square test?

3. How does the answer to Question 1 depend on which IRT model is used (i.e., one-, two-, or three-parameter model)?

It is important to note that we have chosen to use a real test for purposes of illustration only. The location and the width of the range of θs for the closed interval of interest, of course, vary from test to test. We do not expect that the specific nature of our results will generalize to a different test; an easier or a harder test, for example, would likely generate different results. It is hoped, however, that illustrative studies such as this will make it easier for test development practitioners to conceptualize DIF not just as a property of an item overall, but as a value that varies across the ability range for each item. By concerning themselves not so much with whether an item is potentially biased, but instead with where it may be biased, test developers can make more informed decisions about test construction.

Methods

The instrument used was a criterion-referenced test of reading comprehension and mathematics for third graders. It was administered as part of a statewide testing program, and passing it was a requirement for promotion to fourth grade. Not surprisingly, then, the test consisted largely of very easy items. Using the one-parameter model, the cutoff scores for passing the test were equivalent to $\theta = -1.82$ for reading and $\theta = -1.81$ for mathematics. (These scores were, of course, determined by a state standard-setting procedure and have no mathematical bearing on the results of this study; we mention them only because we will later focus the discussion of our results on the ability interval containing them.) The reading and mathematics subtests consisted of 86 and 85 items, respectively. All analyses in the study were conducted on the reading and mathematics subtests separately.

Random samples of 1000 Black and 1000 White examinees were drawn and reanalyzed for this study. In addition, another random sample of 1000 Whites was drawn for use in establishing a baseline.

Item parameters were estimated using PC-BILOG3 (Mislevy & Bock, 1990). Parameter estimates for Black and White examinees were then put on a common scale using the test characteristic curve method (Stocking & Lord, 1983) as implemented in the program EQUATE (Baker, 1990). DIF measures were then computed using Kim & Cohen's IRTDIF program (1992).

For purposes of comparison, parameter estimation and DIF analyses were carried out using the one-parameter (1P), two-

parameter (2P), and three-parameter (3P) logistic models. The three-parameter model was used for two different analyses. First, the analysis was conducted using a variable guessing parameter. Second, the data were reanalyzed with the guessing parameter fixed at .24, a value that was arrived at using a random sample of 500 Black and 500 White examinees. This latter analysis will subsequently be referred to using the abbreviation 3P-c.

DIF indexes were computed separately for a θ range of (-4, 4), which is the Q0 measure as described earlier, as well as for four more limited intervals: (-4, -2), (-2, 0), (0, 2), and (2, 4), which will be referred to as Q1, Q2, Q3, and Q4, respectively. Notice that our focus was the (-2, 0) interval (i.e., the Q2 measure), which contained the cutoff θ for the test. The DIF measures computed included closed-interval signed and unsigned indexes, Raju's exact signed and unsigned measures, and Lord's chi-square statistic. For the exact area and chi-square indexes, tests of significance were conducted automatically through IRTDIF at the .05 and .01 levels. Since no distribution is available for closed-interval area measures, criteria for significance of these at the .05 and .01 levels were established using a White-vs.-White baseline comparison. See Kim & Cohen (1991) for a description of this method.

A descriptive statistics analysis including frequencies and correlations was conducted to determine the degree to which the different measures yielded similar results. Of greatest relevance for this study were the correlations between the closed-interval

measures on Q0 and closed-interval measures on Q2. We also investigated the extent to which the patterns of correlations among models and among different DIF techniques were affected by the use of limited closed-interval measures.

For each item on the test, two ICCs (one from the Black group and another from the White group) were generated graphically using the program IBIAS (Neel, 1993). For items that were flagged for DIF by the Q2 measure but not by the Q0 measure, or vice-versa, ICCs were examined to see if any patterns emerged that might explain the discrepancies.

Results

For all four models, the Q2 measure and the Q0 measure very often failed to identify the same items as exhibiting significant DIF. There are two types of disagreement and two types of agreement in identifying DIF between the Q0 and Q2 measures:

1. Both the Q0 and Q2 measures show no DIF.
2. The Q0 measure shows DIF, but the Q2 measure does not.
3. The Q2 measure shows DIF, but the Q0 measure does not.
4. Both the Q0 and Q2 measures show DIF.

Our interests are the last three types, especially the disagreement described in 2 and 3. We shall call the last three types, "Type A", "Type B", and "Type C" for 2, 3, and 4, respectively. The Type A and B items have serious consequences in decision-making processes. The Type A items would be considered potentially biased by a conventional DIF analysis (and possibly discarded), but the DIF might not be serious at the critical region of the ability

continuum which includes the cutoff θ. On the other hand, the Type B items would not be considered potentially biased by a conventional DIF analysis, but there may in fact be serious DIF at the critical region. See Table 1 for a summary of the definitions of Type A, B, and C items.

Insert Table 1 about here

The numbers of items falling into each of these three categories are summarized by model (1P, 2P, 3P, 3P-c) in Table 2. These results were obtained using the .05 level of significance.

Insert Table 2 about here

Of particular interest are Type A and Type B items, since each of these items could be classified either as biased or as non-biased depending on which DIF measure was used.

Examination of the ICC graphs for all Type A items revealed that Type A items occurred only when either (a) the item was very easy, with most of the bias coming from Q1 or (b) the item curves crossed in the Q2 region, causing smaller areas in that region. A typical example of each of these is shown graphically in Figure 2.

Insert Figure 2 about here

Most of the Type A items for this test occurred because the first

condition (i.e., very easy item) was met; on the mathematics subtest, all Type A items identified using the 1P, 2P, and 3P models can be explained in this way. Since the data were obtained from a minimal-competency test consisting largely of very easy items, these results are not surprising. Clearly, the second condition (curves crossing) was never met in the 1P model.

Type B items occurred generally when the item difficulty fell within the Q2 range, as illustrated by the graphs presented in Figure 3. Another factor that apparently contributed to the incidence of Type B was the crossing of curves outside the Q2 region.

Insert Figure 3 about here

To investigate which limited interval (Q1, Q2, Q3, or Q4) was most closely associated with the DIF measure across the full ability range (i.e., the Q0 measure), correlations between DIF values obtained for Q0 and each of the four limited intervals are summarized in Table 3.

Insert Table 3 about here

Of the four limited intervals, Q1 had the highest correlation with Q0 across reading and mathematics in every instance except the 1P model for reading and the 3P-c model for reading (signed area

only). Using unsigned area measures, correlations between Q0 and Q1 for the 1P, 2P, 3P, and 3P-c models were .78, .80, .77, and .73 respectively for reading and .85, .93, .91, and .90 for mathematics. Again, this result is not unexpected due to the greater variability in Q1 and the fact that the correlation coefficients reported have not been corrected for restriction of range. Even if a correction formula were used, it is not certain that the results could be generalized to a more difficult test. The lowest correlations were those between Q0 and Q4.

Correlations among results obtained from different DIF indexes are summarized in Table 4. As expected, the use of the limited interval Q2 instead of Q0 decreased the correlation with Raju's exact area indexes, both signed and unsigned. For the 2P model, for example, the correlation of the unsigned indexes decreased from .92 to .41 for reading. One unexpected finding was that the correlation between Lord's chi-square index and the Q2 measure generally increased when Q2 was used instead of Q0. This may be due to the fact that Lord's chi-square is based on a point estimate of means, and the use of Q2 cuts off extreme values that affect the DIF index when Q0 is used.

Insert Table 4 about here

Patterns of inter-model agreement for Q2 vs. Q0 can be compared using simple counts of DIF items as given in Table 2. The occurrence of Type A or Type B items appeared to be more pronounced

in relation to Type C items (i.e., agreement of DIF between Q0 and Q2) with higher-parameter models (i.e., 2P or higher), suggesting that the decision as to whether to use the Q0 measure or the Q2 measure is more crucial when higher models are used for the DIF analysis.

To investigate the inter-model agreement within each measure, Q0 or Q2, correlations among different models were calculated for the Q0 measure and the Q2 measure separately (see Table 5). Correlations between models were consistently higher using Q2 than using Q0 for the reading subtest. Using unsigned areas, for example, correlations between the 2P model and the 3P and 3P-c models increased from .87 to .97 and .92 to .97, respectively. Interestingly, this trend is less obvious for the mathematics subtest data.

Insert Table 5 about here

Discussion

The current study illustrates that, at least for this particular test, different items would have been flagged as potentially biased if a limited closed interval had been used, which means that the test developers who selected the items for this test would presumably have made some different decisions. As an example, consider the two items whose curves are shown in Figures 2 and 3 (see Item 30 in Figure 2 vs. Item 28 in Figure 3). Using a DIF index computed over Q0, a test developer would most

likely discard Item 30 and keep Item 28, since the DIF value for Item 30 is almost twice as large (.48 as compared to .25). If, on the other hand, the DIF index were computed over Q2 (the region containing the cutoff score), Item 30 would show a DIF value of only .07, compared to .16 for item 28; the test developer would presumably choose Item 30 over Item 28. Even without knowing the numerical DIF values, it is clear from visually inspecting the curves that the potential bias of Item 28 lies in a range that is critical for this particular test. The curves for Item 30, on the other hand, are divergent mostly in Q1, i.e., for examinees whose scores are substantially lower than the cutoff point.

If situations like the one described above are common, it could be that developers of criterion-referenced tests who use traditional DIF measures based on the entire ability scale are discarding many usable items which exhibit significant DIF values only in a noncritical area of the ability scale. It is also likely that some items whose overall DIF values are not significant may exhibit considerable DIF in precisely that segment of the ability range where it is most crucial to distinguish between competent and incompetent examinees. For this test, such items would have included all items designated as Type B. In the worst-case scenario, pass-fail decisions could be made based on items that are potentially biased for precisely those examinees whose scores are borderline.

The current study has two limitations. First, as previously noted, the specific pattern of results cannot be expected to

generalize to a different test, especially to one composed largely of more difficult items. Second, it is unfortunate that no known distribution exists for the DIF indexes based on closed-interval measures. We maintain, however, that the study is nonetheless valuable in that it suggests, and illustrates, a new way of looking at bias for criterion-referenced tests. While we would not be justified in concluding from this study that limited closed-interval measures are better than traditional DIF measures, it is clear that this possibility should be considered, and that further study in this area would be valuable.

Conceptualizing bias in the way suggested here would seem to give test development practitioners more precise control in test construction. If this new method were to be used in practice, test developers would need to determine the location and the width of the limited θ range, which would depend on the nature and the purpose of the test. Thus, a more active role by practitioners is presumed with this new approach. With user-friendly PC programs such as IRTDIF, which calculates closed interval measures, and a graphics program such as IBIAS which displays DIF via ICCs, practitioners can be easily (and actively) involved in the process of DIF analysis. Instead of allowing one numerical index to determine for them whether an item is or is not potentially biased, practitioners equipped with multiple ways of looking at bias can make more informed and possibly more sound choices about which items should be included in a test.

References

Baker, F. B. (1990). EQUATE: Computer program for equating two matrices in item response theory. Madison, WI: University of Wisconsin, Laboratory of Experimental Design.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

Kim, S, & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. Applied Psychological Measurement, 15(3), 269-278.

Kim, S, & Cohen, A. S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. Applied Psychological Measurement, 16(2), 158.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Neel, J. H. (1993). IBIAS [Computer program]. Atlanta, GA: Georgia State University.

Raju, N. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Table 1

Four Types of Agreement and Disagreement on DIF Between Q0 and Q2

		Q0	
		No-DIF	DIF
Q2	No-DIF		Type A
	DIF	Type B	Type C

Table 2

Number of DIF items by Type (A, B, or C) and Model

	1P		2P		3P		3P-c	
	S	U	S	U	S	U	S	U
Reading								
Type A	2	3	1	2	0	0	2	2
Type B	4	4	11	13	14	12	11	15
Type C	25	24	3	1	4	3	5	2
Math								
Type A	9	10	3	4	5	2	5	5
Type B	4	1	6	7	6	8	4	7
Type C	18	17	1	2	3	2	3	1

S = Signed Area U = Unsigned Area

Table 3

Correlations Between the Q0 Measure and Limited-Interval Measures (Q1, Q2, Q3, or Q4)

	Signed Area				Unsigned Area			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Reading								
1P	.78	.92	.59	.54	.78	.92	.59	.54
2P	.69	.69	.09	.01	.80	.65	.42	.37
3P	.71	.67	.21	.05	.77	.75	.44	.28
3P-c	.62	.76	.03	-.05	.73	.70	.44	.43
Math								
1P	.85	.68	.36	.31	.85	.69	.36	.31
2P	.87	.61	-.10	-.10	.93	.62	.27	.21
3P	.84	.65	.02	-.10	.91	.73	.34	.18
3P-c	.84	.69	-.15	-.15	.90	.69	.23	.20

Table 4

Correlations Between the Q0 Measure and Other DIF Indexes and Correlations Between the Q2 Measure and Other DIF Indexes

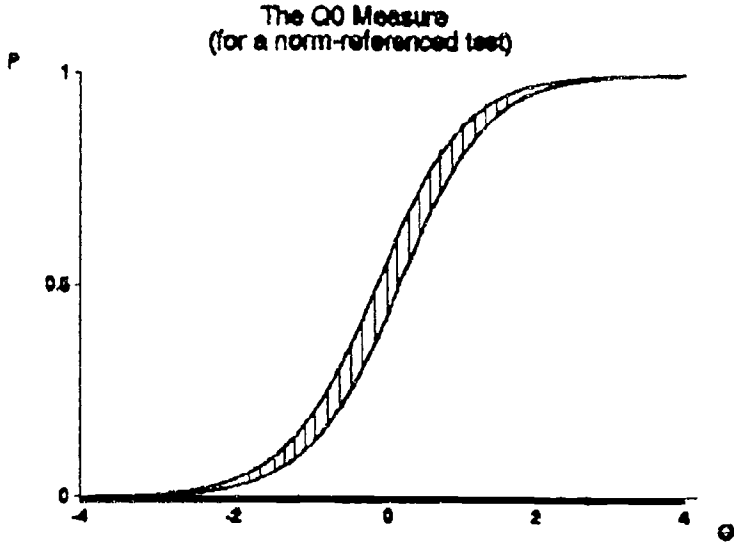
	The Q0 Measure		The Q2 Measure	
		Reading		
Signed				
	Lord's chi-sq.	ESA	Lord's chi-sq.	ESA
1P	.91	.99	.94	.88
2P	.50	.90	.83	.41
3P	.12	N/A	.04	N/A
3P-c	.57	.91	.78	.51
Unsigned				
	Lord's chi-sq.	EUA	Lord's chi-sq.	EUA
1P	.91	.99	.94	.88
2P	.58	.92	.87	.41
3P	.32	N/A	.08	N/A
3P-c	.61	.92	.82	.46
		Math		
Signed				
	Lord's chi-sq.	ESA	Lord's chi-sq.	ESA
1P	.95	.88	.73	.41
2P	.51	.76	.62	.21
3P	.07	N/A	.32	N/A
3P-c	.52	.79	.61	.31
Unsigned				
	Lord's chi-sq.	EUA	Lord's chi-sq.	EUA
1P	.95	.88	.73	.41
2P	.70	.76	.81	.20
3P	.39	N/A	.11	N/A
3P-c	.72	.78	.80	.28

EUA = Raju's exact unsigned area
ESA = Raju's exact signed area

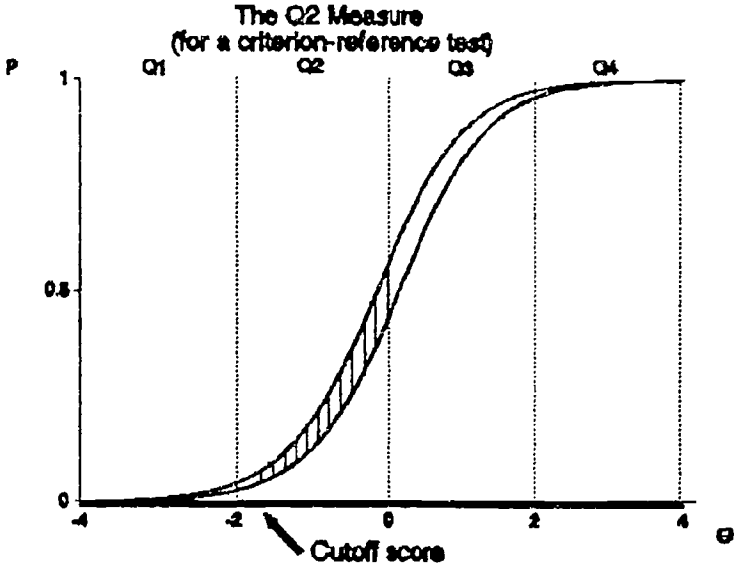
Table 5

Correlations Among Different Models for the Q0 Measure and the Q2 Measure

	The Q0 measure				The Q2 Measure			
Reading								
Signed								
	1P	2P	3P	3P-c	1P	2P	3P	3P-c
1P		.74	.82	.83		.87	.88	.88
2P			.92	.97			.99	.99
3P				.91				1.00
Unsigned								
	1P	2P	3P	3P-c	1P	2P	3P	3P-c
1P		.38	.54	.48		.74	.77	.76
2P			.87	.92			.97	.97
3P				.87				.99
Math								
Signed								
	1P	2P	3P	3P-c	1P	2P	3P	3P-c
1P		.58	.70	.70		.64	.69	.72
2P			.96	.98			.99	.98
3P				.96				1.00
Unsigned								
	1P	2P	3P	3P-c	1P	2P	3P	3P-c
1P		.38	.42	.47		.43	.42	.47
2P			.93	.95			.97	.96
3P				.94				.99

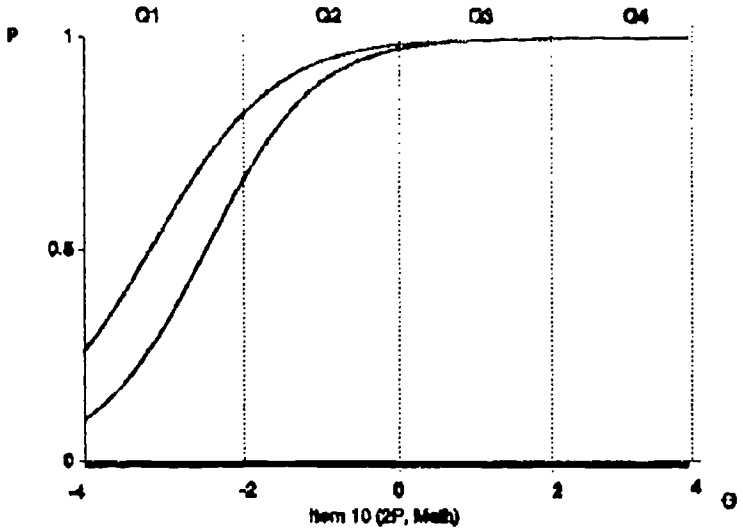


(a)

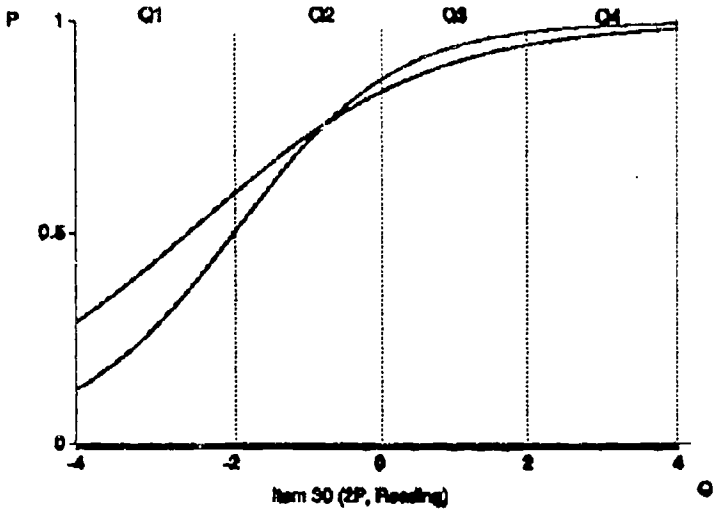


(b)

FIGURE 1. Area based on the full ability range (a) and area based on a limited closed interval (b)

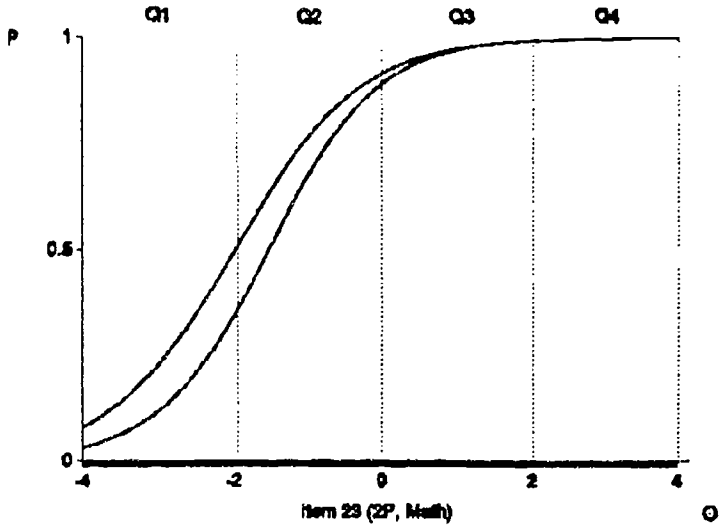


(a)

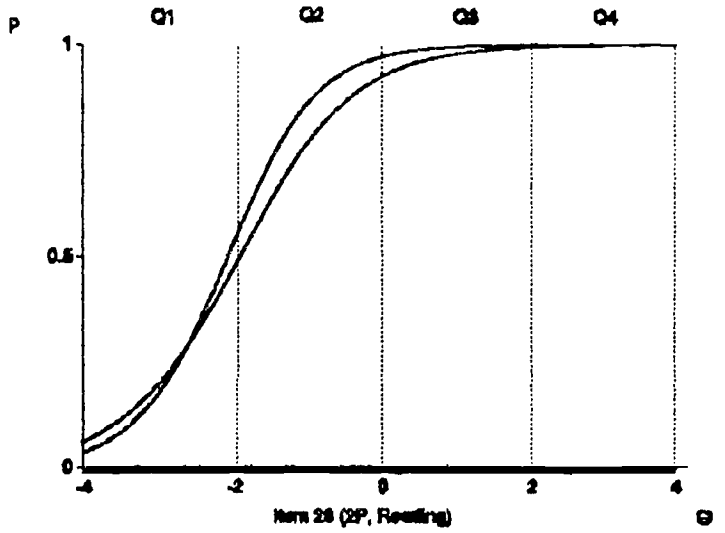


(b)

FIGURE 2. Examples of Type A items



(a)



(b)

FIGURE 3. Examples of Type B items