DOCUMENT RESUME

ED 272 578 TM 860 494

AUTHOR Ackerman, Terry A.
TITLE An Examination of the Relationship between Normalized Residuals and Item Information.
PUB DATE Apr 86
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Computer Simulation; Educational Research; Error of Measurement; *Goodness of Fit; *Item Analysis; Item Banks; Mathematics Tests; Statistical Studies
IDENTIFIERS Normalizing Transformation; *Residuals (Statistics)

ABSTRACT

The purpose of this paper is to present two new alternative methods to the current goodness of fit methodology. With the increase use of computerized adaptive test (CAT), the ability to determine the accuracy of calibrated item parameter estimates is paramount. The first method applies a normalizing transformation to the logistic residuals to make them more interpretable. The second method translates residuals directly into a loss of information statistic. Both methods require a CAT simulation to accurately assess the ability range over which an item would most likely be chosen. Results suggest that the lack of fit in the logistic regression should not be a major concern in developing a CAT item pool. Suggestions for further research are made. (Author)

An Examination of the Relationship between Normalized Residuals and Item Information

Terry A. Ackerman

The American College Testing Program

Paper presented at the 1986 AERA Annual Meeting
San Francisco, CA, April 19, 1986

Running Head: Examination of Logistic Residuals

Abstract

The purpose of this paper is to present two new alternative methods to the current goodness of fit methodology. With the increase use of computerized adaptive test, (CAT) the ability to determine the accuracy of calibrated item parameter estimates is paramount. The first method applies a normalizing transformation to the logistic residuals to make them more interpretable. The second method translates residuals directly into a loss of information statistic. Both methods require a CAT simulation to accurately assess the ability range over which an item would most likely be chosen. Results suggest that the lack of fit in the logistic regression should not be a major concern in developing a CAT item pool. Suggestions for further research are made.

An Examination of the Relationship between Normalized Residual and Item Information

Computerized adaptive testing (CAT) has proven to be a powerful alternative to traditional pencil and paper test administration (Green, Bock, Humphreys, Linn, & Reckase, 1984). In the CAT process the computer selects and administers only items which yield the most information about an examinee's current estimate of ability. Thus the length of a test and the administration time can be shortened considerably without loss of information. Specifically, after responding to an item, an examinee's ability is estimated and an item which yields the most information at that ability is subsequently selected and administered. Items which form CAT item pools are usually selected from previous pencil and paper exams that have been administered and calibrated. A major concern which arises from this process, since only a fraction of the items will be used to estimate any one person's ability, is the accuracy of the obtained parameter estimates. If item parameter estimates do not accurately reflect the true parameters, the item information and ultimately the CAT process will be inaccurate. Several methods have been proposed to assess the goodness of fit of IRT parameter estimates to item response data (Bock, 1972; Yen, 1981; Wright & Mead, 1977; and Bishop, Fienberg, & Holland, 1975). However, no one method appears to be significantly better than the others (McKinley & Mills, 1985).

It is the purpose of this paper to briefly discuss one of the shortcomings of the current goodness of fit methodology and present the findings of two new alternative methods which can be used as part of the selection criteria for CAT item pools. The first alternative is the use of a normalizing transformation proposed by Cox and Snell (1968) on the logistic residuals so that their size and direction can be more readily interpreted.

The second alternative translates logistic residuals directly into a "misinformation" value. Both of these alternatives rely upon the simulation of the CAT process to accurately assess the ability range over which an item would be most likely chosen.

Theoretical Background

Several of the χ^2 goodness of fit statistics used in the research cited above have questionable validity. Hambleton and Swaminathan (1985) discuss some of the problems associated with χ^2 goodness of fit tests including determining the appropriate degrees of freedom, the asymptotically distributed nature of the test criteria and the effect sample size has on the power of the statistic.

Another problem is that the overall χ^2 value may not be indicative of an item's usefulness to the CAT pool. For example, consider Tables 1 and 2 which represent a goodness of fit analysis of two items selected from the ACT Assessment Program's math subtests.

Insert Tables 1 and 2 about here

Table 1 shows Bock's (1972) goodness of fit analysis for item 14, whose parameter estimates are: $\hat{a} = 0.790$, $\hat{b} = 0.604$, $\hat{c} = .200$. To compute the goodness-of-fit statistic the subjects were arranged in increasing order by ability estimate and then divided equally into ten cells. The overall χ^2 goodness of fit value for this item is 20.419 with $p = .005$. Using just this information, one might reject

the item for inclusion into a CAT pool because the estimated ICC appears not to fit the response data very well. However, an important criterion which needs to be considered is the ability range for which this item would be most likely selected. If it was determined (e.g., through a CAT simulation) that item 14 would be most likely chosen in the θ range from -1.27 to .27, then the item should be considered for the pool. That is, the estimated ICC is quite accurately describing the response data as can be seen by the low χ^2 values for those deciles whose Min. θ and Max. θ are in this range.

Table 2 illustrates the opposite situation. The overall χ^2 value for item 5 is 9.678 with $p = 0.208$. The low χ^2 value would suggest a good fit of the model and parameter estimates to the item responses. However, if it was determined that the item would probably be selected in the range from -2.99 to -.83 then the inclusion of this item into the pool should be questioned, since it is in this range that the estimated ICC provides the greatest lack of fit.

Thus the point to be made is that the overall goodness of fit statistic can be easily misinterpreted. A better understanding of the accuracy of the parameter estimates can be achieved by examining the fit of the model in the ability range in which an item is most likely to be chosen.

Experiment 1

Method

Subjects. In the first experiment the normalizing transformation by Cox and Snell (1968) was evaluated using simulated data. One thousand subjects were randomly generated from a N(0,1) distribution. The mean ability of the generated subjects was .01 with a standard deviation of .97.

Materials

An adaptive math test was simulated for each subject. The item pool was composed of 100 items selected from the Mathematics Usage subtests of Forms 26A, 26B, and 26C of the ACT Assessment Program. The items were calibrated using a three parameter logistic (3PL) IRT model by the computer program LOGIST 5 (Wood, Wingersky, Lord, 1982). The sizes of the calibration samples were 2733, 2767, and 2825, respectively. The parameter estimates for the items from 26B and 26C were rescaled to the scale defined by 26A.

In the CAT process an item was selected if it provided the maximum information at the current estimated theta level. The first item "administered to an examinee" in each of the CAT tests was selected based upon the generated ability for that examinee. The testing was terminated if the selected item had an information value $\leq .3$, or if the maximum number of items (20) was administered.

Procedure

Although helpful, the χ^2 goodness of fit measures represent rather gross assessments of how well the data is actually fit by the estimated item parameters. It is usually computed based on deciles of the theta scale whose expected value is determined using the mean or median theta value. "Underfit" or "overfit" of the estimated ICC cannot be determined from the χ^2 value. Such weaknesses can be overcome using a transformation developed by Cox and Snell (1968). According to Cox and Snell, transformed differences between observed

and expected values can be normalized for each estimated theta level. Logistic regression residuals can be transformed to normality according to the following formula

$$RES_i = \frac{n_i[\phi(y_i/n_i) - \phi\{P_i - 1/6(1-2P_i/n_i)\}}{P_i^{1/6}(1-P_i)^{1/6}}$$

RES_i = normalized residual at θ_i

where ϕ ( ) is the incomplete beta function, $I_u(2/3,2/3)$

y_i is the number of correct responses to the item for θ_i

n_i is the number of examinees with θ_i

p_i is the probability of a correct response at θ_i using the 3PL IRT model and the $\hat{a}$, $\hat{b}$, and $\hat{c}$ parameter estimates

Cox and Snell (1968) suggest that the obtained set of residuals is essentially normally distributed for n_i as small as 5 and p_i = .04.

This method has several advantages over the χ^2 goodness of fit measures. First, a normalized residual may be obtained for each estimated ability level (provided $n_i \geq 5$). This eliminates the concern over the optimum number of categories to use in grouping ability levels, or condensing observed and expected value information in a single value per decile.

Secondly, residuals are signed numbers, thus a positive residual would imply the model underestimates the observed proportion correct, while a

negative residual would imply a model overestimates. Thirdly, because the residuals are normalized, their importance can be readily interpreted as z scores.

For each generated examinee the item and current ability estimate were noted for each simulated test. Following the simulation the ability range over which each item was selected was then calculated. Residuals, using the respective calibration response data and item parameter estimates, were then normalized for each theta having $n \geq 5$ in the selected ability range.

Results

Results of the simulation are reported in Table 3. The average length of a simulated test was 17 items. However, of the 100 items in the item pool only 48 items were selected. The parameter estimates and the χ^2 goodness of fit values (using the original calibration samples) for these items are reported in Table 4.

Insert Tables 3 and 4 about here

The position and number of times each item was selected is reported in Table 5. By examining this type of table one can gain a better understanding of the item selection process using specified items. For example it can be seen that the number of items selected increases gradually from 10 in the first position to 48 in the twentieth position.

The minimum and maximum theta values for which each of the 48 chosen items were selected and the number of times each item was selected are shown

in Table 6. The average theta range for each of the 48 items is 1.65. The number of times each item was "administered" ranged from 6 (item 38) to 798 (item 23).

The average theta range for items selected in the first position was .730. The size of the average theta range for the twenty positions varies from .518 for items selected in the second and nineteenth positions to .933 for items selected in the seventh position.

Insert Tables 5 and 6 about here

No significant residuals ($p \leq .05$, $NRES \geq 1.96$) were found at any of the theta values for any of the selected items. The average of the absolute value of the normalized residuals $|RES|$ for each item are reported in the last column of Table 2. This average ranges from .308 to .719 with the majority lying between .31 and .35.

Discussion

The "lack of sensitivity" of the normalizing process to detect significant residuals is partly due to the average n per theta. By rearranging formula (1), it can be shown that the Cox and Snell (1968) transformation is dependent on sample size. That is, the larger the sample size per theta, the smaller the confidence band becomes around the estimated ICC. This is illustrated graphically in Figure 3. In Figure 3 the 95% confidence band around on the ICC for item 1 ($a = .89$, $b = -.99$, $c = .16$) for

n = 10, 50 and 100 are plotted. These confidence bands were calculated by rearranging (1) and solving for the observed p, y_i/n_i, when n and RES_i are specified. The formula to compute the upper 95% confidence limit when n = 10 is given as,

$$\frac{y_i}{n_i} = \left\{ \phi^{-1} \left[\frac{1.96\ P_i^{1/6}(1-P_i)^{1/6}}{10} + \phi\{P_i - 1/6(1 - 2P_i/10)\} \right] \right\} \qquad (2)$$

Insert Figure 1 about here

It can be seen that for the 15 residuals within the targeted ability range for item 1, none was beyond the 95% confidence band until n = 50. This is considerably above the average of 8 subjects per theta used in this study. If the sample size was at least this large two thirds of these residuals would have been significant at $p < .05$. (However, it is questionable that such residuals would exist with such a large sample size.)

The correlation between the χ^2 goodness of fit value and $|\overline{RES}|$ was -.177. This lack of linear relationship is probably due, in part, to the sensitivity of the normalized residual analysis, which appears to detect a great deal of "noise" common to any regression analysis. However, the advantages thought to be gained by this technique may not be that helpful unless a larger calibration sample is used (i.e. large enough to yield an n per theta $\geq$ 50).

Experiment 2

Since item selection within the adaptive testing process is directly related to the amount of information each item provides, it was decided in the second part of this study to investigate how much information would be lost if the estimated parameters were changed to fit the observed calibration data exactly within the targeted theta range.

Method

Subjects and Materials

The simulated examinee and adaptive test results used in Experiment 1 were also used in Experiment 2. That is, the 1000 generated subjects, the results of their simulated adaptive tests, and the calibration data were reanalyzed in the second part of this study.

Design and Procedure

To investigate the amount of mis-information which occurs from the lack of fit. the following statistic was derived

$$MIS_j = \frac{\sum_{i=1}^{k} I_{b_j} - I_{b_i}}{k}$$

where MIS_j = Mis-information statistic for item j

I_{b_j} = Item information value for item j using the originally calibrated difficulty parameter "b_j"

$I_{\hat{b}_i}$ = Item information value using the adjusted $\hat{b}_i$ value for theta "i".

k = # of thetas ($n \geq 5$) in the ability range in which item j was selected

The following steps were performed to calculate the MIS_j statistic for each item:

Step 1: For each of the thetas ($n \geq 5$) in the ability range for which an item was selected in the CAT simulation, the observed proportion correct in the calibration sample was computed.

Step 2: Using the observed proportion correct, a new difficulty $\hat{b}_i$ was calculated, for each $\hat{\theta}_i$. This $\hat{b}_i$ is what the difficulty would have to be if the observed p for the estimated ICC were to become the expected p. That is, $\hat{b}_i$ is selected so that the new ICC would pass through the observed p value at the given $\hat{\theta}_i$. An assumption is made that the $\hat{a}$ and $\hat{c}$ parameters would remain as originally estimated. If the observed proportion correct $\leq \hat{c}$, then the largest displaced $\hat{b}_i$ is used since obviously no new ICC could be created. This process is represented graphically by the dotted lines in Figure 2 which denote the new ICCs passing through the residuals $\geq \hat{c}$ for item 1.

Insert Figure 2 about here

Step 3: For each theta in the targeted range the information function using $\hat{b}_j$ and $\hat{b}_i$ was determined. The difference between these values at their respective theta levels were then calculated. (Note: the difference between the information functions can be either positive or negative.)

$$\text{If } \hat{\theta}_i - \hat{b}_j < \hat{\theta}_i - \hat{b}_i \text{ then } I_{b_j} - I_{b_i} < 0.$$

$$\text{If } \hat{\theta}_i - \hat{b}_j < \hat{\theta}_i - \hat{b}_i \text{ then } I_{b_j} - I_{b_i} > 0.)$$

This is because the information function will be a maximum near the point $\hat{\theta} \equiv \hat{b}$. (See Lord 1980, p. 152 for the exact θ where the maximum of a 3PL model occurs.)

A graphical representation of this analysis can be seen in Figure 3. The original information function is represented by the dark thick curve, while the two adjusted ones are represented by chain dotted curves. In one instance ($\hat{b}_i = -.45$, $\hat{b}_j = -.99$, $\hat{\theta}_i = -1.05$) the difference can be seen to be positive (dashed line) and in the other ($\hat{b}_i = -1.46$, $\hat{b}_j = -.99$, $\hat{\theta}_i = 1.41$) negative (dotted line).

Insert Figure 3 about here

Step 4: The mean average of the absolute value of the difference, $I_{b_j} - I_{b_i}$, was computed over all of the thetas in the targeted range.

To evaluate this statistic the following ratio was formed.

$$AMIR_j = \frac{\bar{I}_{s_j} - MIS_j}{\bar{I}_{s_j}}$$

where $AMIR_j$ = average mis-information ratio for item j

$\bar{I}_{s_j}$ = the average item information value provided by item j in the simulated CAT

The AMIR ratio is formed in the following manner. The item information value for each estimated theta was calculated for each item and averaged over the number of times the item was selected in the simulated CAT. The mis-information value calculated using the calibration sample was then subtracted from the average information provided when the item was selected in the simulated CAT. The difference was then divided by the average item information value provided in the simulated CAT. Notice that the $AMIR_j$ ratio will only be negative if an item provides more average mis-information than average information in the theta range in which the item is selected. Thus it is believed that if the $AMIR_j < 0$ the lack of fit of the 3PL model to the item response data provides sufficient mis-information that an item should probably not be included in a CAT item pool.

Results and Discussion

The MIS_j values computed using the ability range provided by the simulated adaptive tests and the 3PL parameter estimates and item response from the calibration samples are shown in Table 7.

Insert Table 7 about here

The range of the MIS_j values for the 48 selected items is .12 to .56 with an average of .28. The AMIR ratio for each selected item is also shown in Table 7. The average of the AMIR ratios was .53 with a standard deviation of .36.

Only one item, item 45, has a $AMIR_j$ ratio less than one. However, this item was selected in a range from -4.05 to -1.43, and only four thetas ($n \geq 5$) could be found in this range in the calibration sample. Item 45 was the easiest item ($\hat{b} = -1.552$) out of the 100 items in the pool, and yet it was providing more "mis-information" than information in the selected range. Thus in this case the negative AMIR value could be interpreted as an index describing the item pool, suggesting a need for more easy items. An important aspect which needs to be considered is accurately specifying the targeted test population for the CAT simulation. That is, if it was expected that only high ability students would be taking the CAT then there would probably be no concern for adding more low difficulty items.

General Discussion

The best way to understand if the items collected for a CAT item pool provide effective measurement in the targeted ability range is to simulate the adaptive testing process. Simulation enables one to determine the range over which an item is most likely to be chosen, and thus provide better interpretation of χ^2 goodness of fit analyses.

The method of selection used in the simulated adaptive test for this study was to select the item which provided the most information at the current estimated ability level, however, other methods do exist (see Hulin, Drasgow and Parsons 1983, pp. 226-230.)

Hulin, Drasgow and Parsons (1983) reported similar findings about the number of items selected in the CAT process. Their results, using the maximum information method, revealed that only 119 items out of a 260 item pool were ever selected. These findings when viewed in concert with the results of this study, would suggest that if the criteria for item selection in the CAT process is to select the item which provides the maximum information, over half of the items may never be used.

If one chooses this method of item selection two concerns arise. First, the items which are selected (e.g., in a simulation) need to be checked for the degree of mis-information each provides due to lack of logistic fit. Second, how shall those items not selected be evaluated? One possible solution would be to avoid this problem by restricting (e.g., for security reasons) the number of times an item can be administered. For example, in the present study of the 48 items chosen in the CAT simulation each was selected on an average of 355 times! The minimum number of times each item could be selected so that all 100 items were used equally would be 10 times. However,

the standard error of the ability estimates would be greatly disproportionate between the first examinees and the last examinees due to lack of informative items remaining in the pool. Thus, if a selection restriction is placed upon the items in the pool, it is necessary to have a large enough item pool to provide accurate ability estimates for the entire test population over the targeted range.

The normalized residual transformation although more directly interpretable than a χ^2 goodness of fit test appears to be an infeasible approach because of the large sample sizes needed. Such large samples might make the cost of the calibration process prohibitive.

The results obtained using the $AMIR_j$ ratio suggest that the lack of fit in the logistic regression process should not be a major concern in the selection process for a CAT item pool. Of the 48 items selected in the CAT simulation, 10 would have been rejected outright if the selection criteria was a χ^2 goodness of fit value whose $p \leq .05$. Using the AMIR ratio only one item was flagged, and this in part, was due to the lack of very easy items in the pool, rather than a faulty item. These results are promising when one realizes the time, effort and expense put into item development.

The correlation between the AMIR ratio and the χ^2 goodness of fit value for the 48 items was only .131, suggesting little linear relationship between the two. Normalized residuals, however, do correlate quite highly with the AMIR ratios, $r = -.776$.

More research needs to be conducted to validate the concerns and new approaches presented in this paper. Mis-information analyses needs to be conducted using other methods of item selection. Hopefully the problems which plague goodness of fit analyses may prove to be unwarranted.

References

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: The MIT Press.

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

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. Journal of the Royal Statistical Society, 30, 248-75, Series B.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Evaluation plan for the computerized adaptive vocational aptitude battery (MPL TN 85-1). San Diego: Navy Personnel Research and Development Center.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory. Boston, MA: Kluwer Nijhoff Publishing.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

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

McKinley, R. L. & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Wood, R. L., Wyngersky, M. S., & Lord, F. M. (1976). LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: ETS.

Wright, B. D. & Mead, R. J. (1977) BICAL: Calibrating items and scales with the Rasch model (Research Memorandum No. 23). Chicago, IL: University of Chicago, Statistical Laboratory, Department of Education.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Table 1. Bock's goodness of Fit Analysis for Item 14

Logist Parameter Estimates: $\hat{a} = 0.790$ $\hat{b} = 0.604$ $\hat{c} = 0.200$

Cell	Observed P	Expected P	Cell N	χ^2	Min Theta	Max Theta	Median Theta
1	.293	.233	263.	5.187	-2.99	-1.27	-1.73
2	.305	.281	262.	0.756	-1.27	-0.83	-1.02
3	.322	.325	264.	0.013	-1.83	-0.48	-0.65
4	.398	.376	264.	0.544	-0.48	-0.23	-0.34
5	.428	.424	264.	0.019	-0.23	0.02	-0.10
6	.477	.479	264.	0.004	0.02	0.27	0.14
7	.449	.548	263.	10.524	0.28	0.57	0.41
8	.643	.623	263.	0.425	0.57	0.83	0.69
9	.725	.704	262.	0.567	0.83	1.21	1.00
10	.867	.832	264.	2.380	1.22	2.96	1.59

Note: Bock's Chi squared goodness of fit total is 20.419 with 7.0 degrees of freedom P = 0.005

Table 2. Bock's goodness of Fit Analysis for Item 5

Logist Parameter Estimates: $\hat{a}$ = 1.020 $\hat{b}$ = 0.511 $\hat{c}$ = 0.160

Cell	Observed P	Expected P	Cell N	χ^2	Min Theta	Max Theta	Median Theta
1	.189	.250	264.	5.239	-2.99	-1.27	-1.73
2	.447	.406	264.	1.865	-1.27	-0.83	-1.02
3	.548	.530	263.	0.342	-0.83	-0.48	-0.65
4	.654	.642	263.	0.171	-0.48	-0.23	-0.34
5	.712	.724	264.	0.175	-0.23	0.02	-0.10
6	.795	.795	264.	0.001	0.02	0.27	0.14
7	.856	.859	263.	0.020	0.28	0.57	0.41
8	.890	.907	264.	0.885	0.57	0.83	0.69
9	.955	.943	265.	0.672	0.83	1.21	1.00
10	.974	.979	266.	0.308	1.22	2.96	1.59

Note: Bock's Chi squared goodness of fit total is 9.678 with 7.0 degrees of freedom P = 0.208

Table 3. Descriptive Statistics of the CAT Simulation

Examinees were generated randomly from a N (0,1) distribution

N = 1000	
Mean θ = .01	Minimum = -4.05
S.D. θ = .97	Maximum = 3.80

Estimated abilities

Mean $\hat{\theta}$ = -.02	Minimum = -4.05	$\overline{S.E.}$ = .412
S.D. $\hat{\theta}$ = 1.35	Maximum = 3.87	

Length of simulated tests

Mean = 17.00	Minimum = 1
S.D. = 6.25	Maximum = 20

TABLE 4. Item parameters and goodness of fit statistics for the items selected in the CAT simulation

ITEM	$\hat{a}$	$\hat{b}$	$\hat{c}$	χ^2	p	RES	ITEM	$\hat{a}$	$\hat{b}$	$\hat{c}$	χ^2	p	RES
1	0.89	-0.39	0.16	8.37	.30	.56	48	1.03	-0.70	0.18	33.15	.00	.33
2	0.93	-1.25	0.16	15.34	.03	.34	49	0.93	-0.53	0.18	18.03	.07	.35
5	1.02	-0.51	0.16	9.66	.21	.32	53	1.06	-0.33	0.18	10.46	.16	.33
8	1.04	0.03	0.16	29.95	.00	.35	54	1.14	0.13	0.18	5.76	.57	.32
9	1.15	0.07	0.16	7.83	.35	.34	56	1.27	-0.41	0.18	6.16	.52	.34
12	0.84	-0.65	0.16	17.26	.02	.35	57	1.06	0.66	0.18	17.69	.01	***
21	1.01	0.41	0.19	6.69	.46	.39	59	1.41	-0.40	0.18	7.71	.36	.32
23	1.43	0.17	0.13	7.44	.38	.39	60	1.05	-0.07	0.18	12.43	.09	.32
25	1.65	0.86	0.19	3.97	.78	.37	63	1.72	0.68	0.24	28.06	.00	.40
26	1.03	0.43	0.13	4.83	.68	.37	67	1.15	0.72	0.13	4.52	.72	.35
28	1.57	0.29	0.24	10.94	.14	.33	68	1.02	1.07	0.12	6.78	.27	.57
29	0.97	0.17	0.09	8.58	.28	.32	70	2.01	1.28	0.24	11.08	.14	.36
30	1.60	0.55	0.23	4.59	.71	.37	73	1.10	1.16	0.20	10.56	.16	.46
32	1.49	0.60	0.09	10.04	.19	.38	74	1.32	0.62	0.12	3.32	.85	.36
33	1.01	0.98	0.25	5.52	.55	.34	75	1.42	1.41	0.26	2.93	.89	.45
35	1.16	0.89	0.16	4.90	.67	.43	76	1.74	1.02	0.25	6.97	.43	.39
36	1.59	0.72	0.16	2.14	.95	.38	77	1.28	1.74	0.15	13.67	.06	.55
37	1.75	0.95	0.14	11.98	.10	***	78	1.10	0.69	0.01	16.27	.02	.36
38	0.93	0.60	0.16	8.82	.27	***	84	0.93	-0.57	0.16	17.34	.02	.34
39	1.75	0.76	0.20	8.21	.31	.39	85	1.29	-0.88	0.16	16.13	.02	.34
40	0.93	1.17	0.09	5.06	.65	.4[illegible]	86	0.95	-1.14	0.16	6.19	.52	.33
41	0.96	-0.72	0.18	13.69	.06	.31	87	0.95	-0.65	0.16	8.29	.31	.32
44	1.04	-1.33	0.18	9.25	.24	.33	90	0.81	-1.02	0.16	14.20	.05	.34
45	0.74	-1.55	0.18	5.01	.66	.72	91	1.08	-0.64	0.16	11.77	.11	.34

*** Note: items for which no residuals (n = 5) were found in the calibration sample in the ability range in which the item was selected

Table 5. Position of Item Selection

	POSITION CHOSEN																			
ITEM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	0	0	0	0	1	4	1	2			2	4	3	5	3	2
2	0	14	104	24	15	0	14	21	13	26	14	16	8	.	10	13	7	15	16	14
5	0	0	0	0	126	48	127	29	29	38	3	29	24	25	8	9	9	17	8	11
8	0	0	0	0	0	36	26	45	36	78	15	121	79	65	44	16	13	16	21	26
9	0	0	0	0	0	0	0	0	0	1	11	13	20	30	93	140	114	65	37	44
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	5	6	5	10
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	39	34
23	43	26	9	284	29	71	56	71	13	54	36	19	17	11	6	15	12	3	17	6
25	72	59	99	33	27	26	40	31	33	12	25	18	20	9	7	8	5	1	2	3
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	48	78	77	117
28	0	0	0	0	0	0	0	35	36	29	67	17	64	85	113	61	28	41	35	24
29	0	0	0	0	0	0	0	0	0	0	0	5	14	39	13	32	63	102	60	67
30	0	0	120	51	109	114	92	76	23	20	18	13	9	6	4	4	6	6	0	4
32	97	72	192	170	73	13	19	5	14	2	13	4	1	3	2	5	5	0	2	2
33	0	0	0	0	0	0	0	0	0	46	82	88	50	35	28	15	20	11	6	9
35	0	0	0	0	0	0	0	0	0	0	0	0	0	11	46	86	62	22	14	7
36	0	0	3	46	151	197	48	43	25	24	16	3	8	9	7	6	0	4	4	4
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	20	6
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
39	167	147	67	40	58	27	44	39	5	6	2	4	8	6	1	3	2	2	4	3
40	0	2	13	12	2	1	0	0	26	2	0	4	37	55	42	19	13	8	12	11
41	0	0	0	0	0	0	0	1	2	23	13	9	11	17	6	11	14	17	20	16
44	36	118	27	75	5	0	10	2	20	19	8	6	4	10	15	4	9	8	11	6
45	2	7	23	41	6	9	0	8	1	1	1	2	1	1	4	0	1	1	1	1
48	0	0	0	0	0	0	18	43	60	18	29	14	11	21	14	11	18	17	13	15
49	0	0	0	0	0	0	0	0	0	0	0	0	0	8	12	5	18	6	11	19
52	0	0	0	0	0	0	0	0	0	6	33	57	53	24	23	18	30	19	28	41
54	0	0	0	0	0	0	0	0	0	0	0	0	5	8	51	40	107	74	83	50
56	0	0	176	28	172	14	54	5	19	24	45	4	11	9	5	9	11	7	5	7
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	12
59	287	259	11	35	13	1	3	8	10	3	10	14	10	7	1	3	8	5	1	2
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	7	21	38	64
63	0	0	0	0	0	37	163	137	54	66	48	21	8	9	5	10	6	7	6	4
67	0	0	0	0	0	0	0	0	0	0	0	0	67	79	100	56	28	32	31	27
68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	32	39	21
70	74	112	38	10	10	3	1	10	9	9	10	11	9	6	9	14	8	6	4	5
73	0	0	0	0	0	0	0	0	0	0	0	0	10	0	1	7	24	23	8	11
74	0	0	0	0	0	0	0	42	277	108	61	53	46	23	13	6	7	15	7	1
75	0	2	11	13	3	0	0	34	19	1	16	9	9	6	7	14	8	6	9	6
76	0	0	0	0	44	51	35	62	46	38	23	13	28	23	20	16	6	7	6	7
77	5	57	51	11	3	1	0	0	0	16	0	2	1	2	1	5	14	8	4	4
78	0	0	0	0	0	0	0	0	8	120	168	189	115	93	24	15	10	7	11	11
84	0	0	0	0	0	0	0	0	0	0	0	11	13	8	6	20	2	15	16	22
85	217	118	24	11	0	123	0	15	2	3	14	4	12	7	14	5	4	6	8	5
86	0	0	0	7	4	0	13	4	24	17	11	10	12	12	13	7	20	12	5	12
87	0	0	0	0	0	0	0	0	0	1	18	10	10	16	18	28	13	18	24	17
90	0	0	0	0	0	0	0	0	0	0	0	1	0	1	4	3	8	3	4	3
91	0	0	0	0	0	78	73	30	21	13	9	21	8	19	15	23	11	19	21	11

Table 6. Minimum and Maximum Values for which Each Selected Item was Chosen

Item	N*	Min $\hat{\theta}$	Max $\hat{\theta}$	Item	N*	Min $\hat{\theta}$	Max $\hat{\theta}$
1	(36)	-1.93	-1.05	48	(299)	-1.67	-0.03
2	(336)	-2.58	-0.20	49	(70)	-1.23	-0.62
5	(540)	-1.21	0.52	53	(332)	-1.27	0.18
8	(639)	-1.04	0.94	54	(418)	-0.59	0.89
9	(568)	-0.53	1.18	56	(605)	-1.05	0.63
12	(38)	-1.42	-0.82	57	(18)	1.81	2.29
21	(79)	0.90	1.23	59	(691)	-1.23	0.80
23	(798)	-0.83	1.77	60	(131)	-0.79	0.08
25	(530)	0.21	1.73	63	(581)	-0.01	2.04
26	(349)	-0.18	1.78	67	(420)	0.19	2.23
28	(635)	-1.20	0.99	68	(160)	1.18	2.15
29	(395)	-1.19	1.09	70	(358)	0.66	2.41
30	(675)	-0.21	1.94	73	(84)	1.25	2.09
32	(714)	-0.27	1.77	74	(659)	-0.33	2.14
33	(392)	0.46	2.02	75	(173)	0.99	3.33
35	(248)	0.70	2.20	76	(425)	0.42	1.82
36	(596)	0.03	1.86	77	(185)	1.10	3.80
37	(32)	1.68	2.25	78	(771)	-0.97	2.05
38	(6)	2.31	2.31	84	(113)	-1.35	-0.57
39	(651)	-0.18	1.86	85	(592)	-1.66	0.42
40	(259)	0.78	3.87	86	(183)	-2.07	-0.55
41	(160)	-1.64	-0.29	87	(173)	-1.56	-0.21
44	(397)	-2.43	-0.27	90	(27)	-1.62	-0.99
45	(111)	-4.05	-1.43	91	(392)	-1.32	0.38

*N = the number of times the item was selected in the CAT simulation.

Table 7. Mis-Information and Average Mis-Information Ratio Values for the Item Selected in the Simulated CAT.

Item	MIS	AMIR	Item	MIS	AMIR
1	.241	.287	48	.149	.712
2	.122	.679	49	.167	.521
5	.134	.725	53	.140	.724
8	.169	.649	54	.233	.724
9	.179	.690	56	.274	.664
12	.123	.597	57	***	***
21	.167	.596	59	.284	.668
23	.369	.548	60	.221	.734
25	.373	.612	63	.379	.478
26	.148	.707	67	.216	.750
28	.458	.339	68	.452	.481
29	.140	.695	70	.512	.286
30	.397	.535	73	.458	.483
32	.442	.421	74	.263	.635
33	.129	.662	75	.553	.242
35	.188	.702	76	.402	.419
36	.393	.572	77	.473	.263
37	***	***	78	.320	.556
38	***	***	84	.299	.636
39	.566	.490	85	.303	.598
40	.142	.670	86	.251	.724
41	.120	.719	87	.257	.713
44	.166	.600	90	.419	.545
45	.211	-1.705	91	.265	.705

*** Denotes items for which MIS and AMIR could not be calculated because there were no thetas in the selected range in the calibration samples.

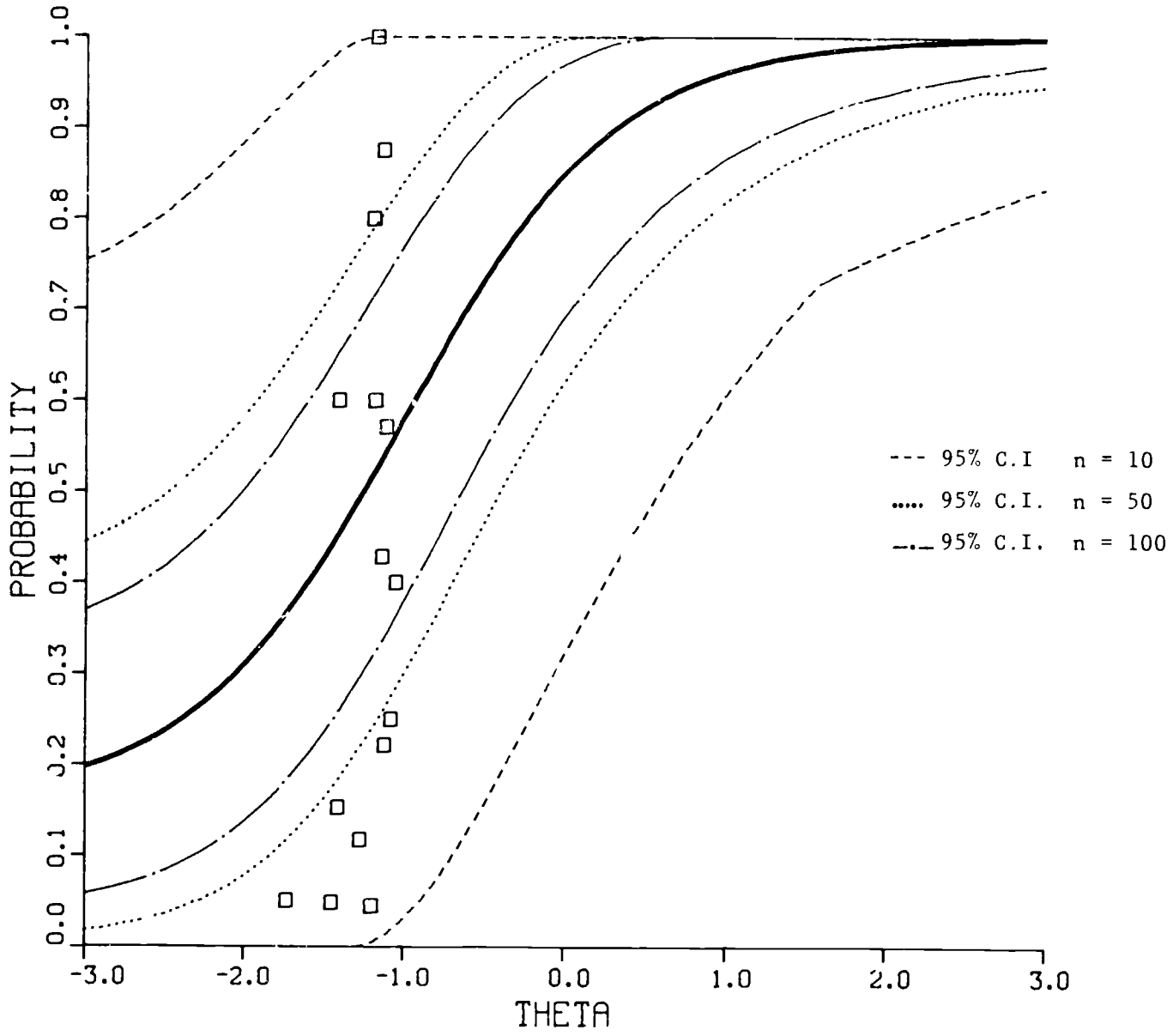

Figure 1. 95% Confidence Intervals Placed About the ICC for Item 1 Using Three Different Sample Sizes

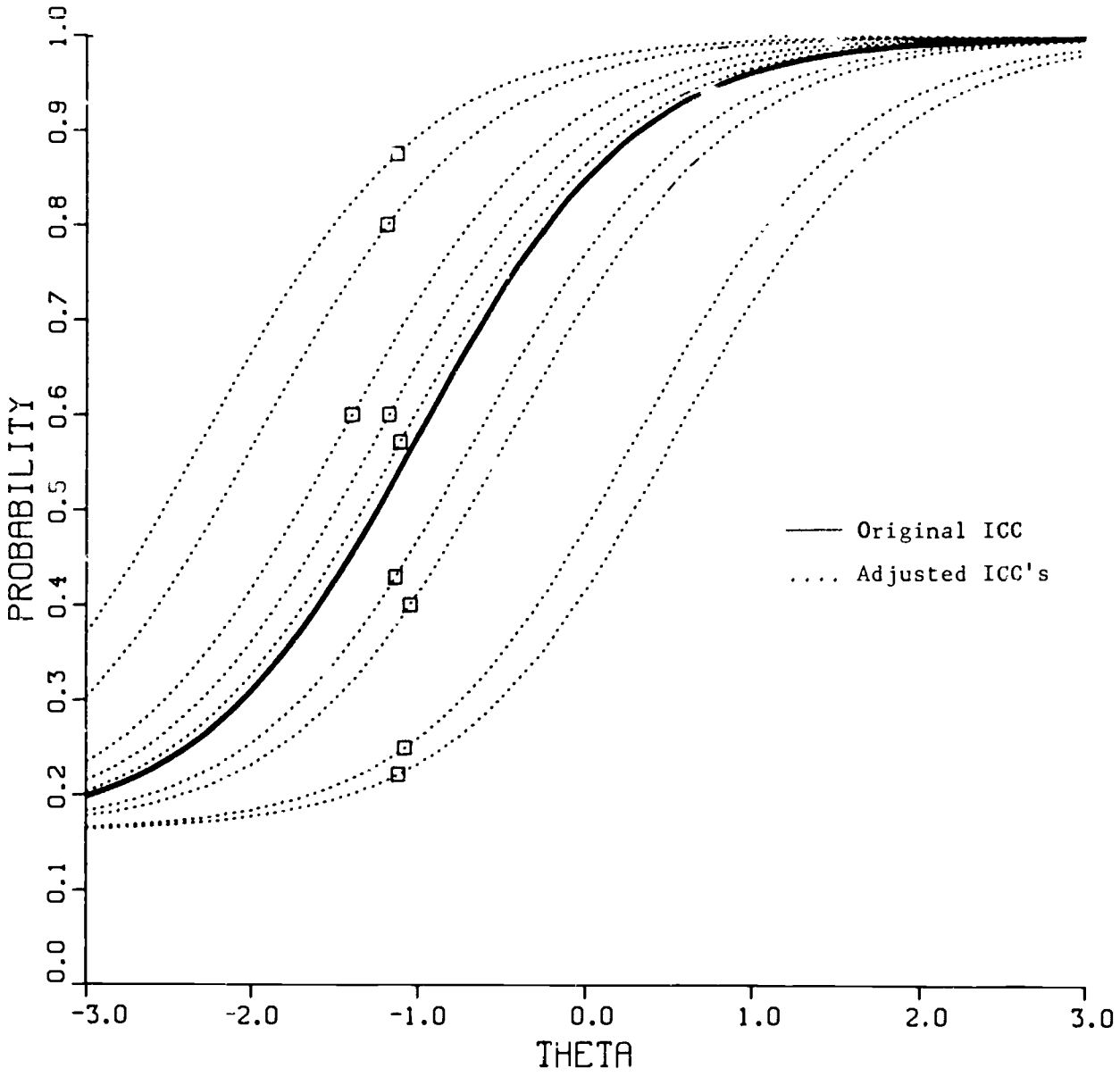

Figure 2. Adjusted ICC's Passing Through the Residuals in the Selected Range for Item 1

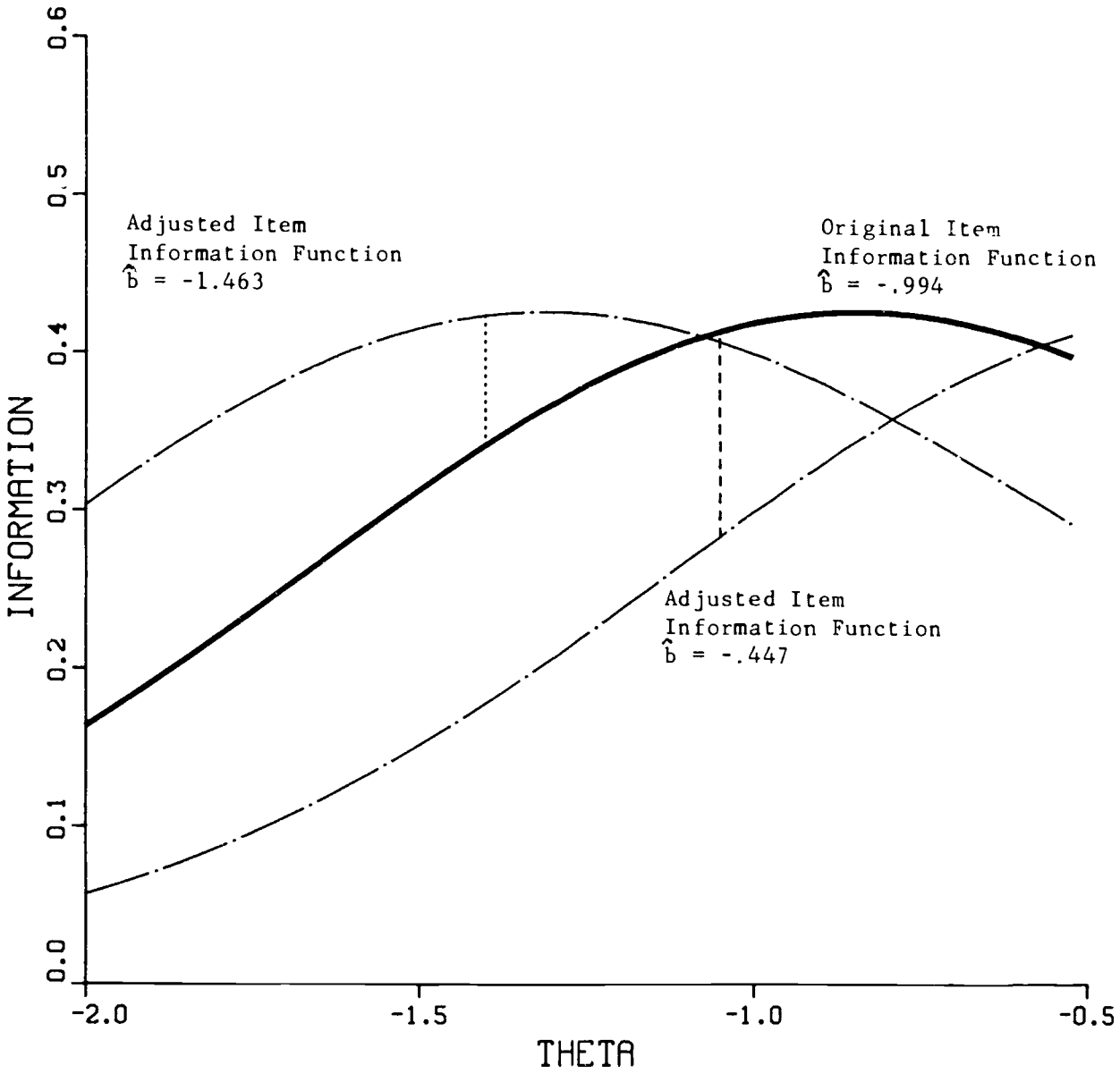

Figure 3. Mis-Information Analysis for Item 1