ACT Research Report Series 88-3

Comparison of Loglinear and Logistic Regression Models for Detecting Changes in Proportions

Judith A. Spray
James E. Carlson

November 1988

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

COMPARISON OF LOGLINEAR AND LOGISTIC REGRESSION MODELS FOR DETECTING CHANGES IN PROPORTIONS

Judith A. Spray

James E. Carlson

ABSTRACT

Two statistical procedures were compared in terms of their ability to detect significant changes in item difficulty and item discrimination from pretest to national test administrations. These two procedures were a loglinear (LL) analysis and a logistic regression (LR) procedure. The results of this study showed that the two procedures were able to identify two items from an ACT Mathematics Usage test showing unstable difficulty and discrimination between two administrations. Both procedures yielded similar results in showing stability in the remaining items, although rank correlations of the results from LL and LR showed some inconsistency when discrimination changes were assessed, probably due to the use of fewer ability (score) categories in LL versus LR.

COMPARISON OF LOGLINEAR AND LOGISTIC REGRESSION MODELS FOR DETECTING CHANGES IN PROPORTIONS

Whenever test items are administered more than one time to samples of examinees that are theoretically from the same population, we should be concerned about sources of variation of proportion-correct values or p from sample to sample. One source of variation could simply be random sampling and if this were the case, differences in p might prove to be almost negligible or at least nonsignificant most of the time. On the other hand, significantly large differences in p might occur between testing administrations if there had been major changes in the item itself, such as the dropping of a popular foil or the rewriting or editing of the item's stem or alternatives.

These significant variations in certain item characteristics or parameters could have major ramifications on test construction. Usual test construction practice is to select items for inclusion in a test by specifying a target distribution for p and then to select a set of items that matches, as closely as possible, that target distribution. If individual item parameters change significantly between test administrations, the actual distributions of these item parameters may no longer match the target distribution.

In some applications, such as computerized adaptive testing, items are administered continually over some period of time. Change in item parameters in such situations has been termed parameter drift (Rentz, 1978). In the present paper, the interest is in situations where the same set of test items is administered twice, to two different samples of examinees. However, the methodology discussed within this paper can easily be extended to more than two administrations of the test items.

The specific situation under consideration is one in which possible changes in p might occur between pretest and national administrations of the items in the ACT Assessment Program. Item statistics obtained from pretest

administrations are used to assemble the final forms of the tests in this program. Obviously, if the item statistics change significantly between pretest and national administrations, then the actual distribution obtained from the national sample could be considerably different from the target specifications.

Although it would be too late to remedy these flawed distributions once the national sample had responded, it would still be helpful to have some method of maintaining a certain level of quality control over the item selection process. Furthermore, if certain types of items tended to be more prone to change than other types, this information would be valuable at the time of test construction.

The purpose of this paper is to compare two techniques or statistical procedures in their ability to detect item parameter changes from pretest to national administrations. The term, "item parameter" is used here and elsewhere to describe proportion-correct values or proportion-correct values at certain achievement levels of the examinees (e.g., test scores). If this problem seems similar to the problem of item bias detection, it is probably because most of the item bias detection procedures that have been proposed in recent years could, in fact, be used in the present circumstances to detect differences in p between test administrations rather than between groups.

We have chosen to compare the results from two procedures, (1) loglinear models and (2) logistic regression analysis, in detecting item parameter differences on different samples at pretest and national administrations. The use of the loglinear (LL) procedure has been suggested as an item bias detection method by a number of authors (Kelderman, 1985; Kok, Mellenbergh, & Van Der Flier, 1985; Marascuilo & Slaughter, 1981; Mellenbergh, 1982; Van Der Flier, Mellenbergh, Ader & Wijn, 1984). To our knowledge, the use of

the logistic regression (LR) procedure for a similar purpose has not been advocated. The two procedures are described in greater detail in the following sections of the paper.

Loglinear Models

Within the context of the present problem, if each item in an n-item test is treated as a separate entity, item responses can be categorized into n, 3-way contingency tables according to three categorical variables: item response (R) with two levels, correct and incorrect; ability (A) with k levels; and test administration (T) with two levels, pretest and national. The frequencies within each cell of a table can be analyzed using a technique called loglinear modeling or analysis (Bishop, Fienberg & Holland, 1975; Fienberg, 1977). The inclusion of the ability variable in the table allows a test of proportion-correct differences from pretest to national administrations at given levels of ability. The detection of these differences has been termed a test of "nonuniform bias" or "conditional methods of bias detection" by Mellenbergh (1982) Van Der Flier et al. (1984) and Kok et al. (1985). However, because an item with good discrimination is expected to have differences in proportion-correct values at different ability levels, the differences (between pretest and national p values at given ability levels) are referred to here as discrimination differences.

The LL procedure consists of modeling the expected value of the log of the cell frequencies or proportions in terms of main, or marginal effects (T = test, R = response, A = ability) and interaction effects (TR = test by response, TA = test by ability, RA = response by ability, and TRA = test by response by ability). The models have an "ANOVA-like" appearance to them except that there is no random error term in the model because there is only one observed frequency per cell; hence, there is no "within-cell" variation.

Another difference between ANOVA and LL models is an artificial function of the way in which these models are typically described in the statistical literature. The LL analysis of cell frequencies is usually conceptualized in terms of comparisons of goodness-of-fit indices of a series of hierarchical models (with and without various interaction terms present). Although ANOVA can be conceptualized in an analogous fashion (e.g., Searle, 1971), the description in applied statistics texts is more often in terms of partitions of variation based on a single model. Each model for the frequency data yields expected counts for each cell and the goodness-of-fit statistic compares the expected counts with the observed counts. The most commonly used statistics for this comparison are the Pearson chi-square statistic and the likelihood-ratio chi-square statistic (abbreviated as G^2). The latter is preferred because differences between values of this statistic for hierarchical models are themselves distributed as chi-square random variables (Bishop, Fienberg & Holland, 1975).

Another difference between ANOVA and LL models is that whereas the parameters of the ANOVA model are naturally additive, those for frequencies from a multi-way contingency table are multiplicative. Hence logarithms of the multiplicative parameters are taken to get an additive, linear model. A cell mean in an ANOVA model, for example, is the sum of parameters associated with each main effect and interaction. An expected cell frequency under a model that assumes independence in a multi-way contingency table, on the other hand, is a product of parameters associated with each main effect. Figure 1 represents the design and the parameters of the LL analyses in this paper. Note that the design pictured in Figure 1 assumes that the number of levels of the ability variable (A) is three (Low, Medium, High).

Referring to this figure, and recalling the product rule for probabilities

of occurrences of independent events, it can be seen that if test, response, and ability are assumed to be independent,

$$p_{ijk} = p_{i..}p_{.j.}p_{..k}, \; i = 1,2; \; j = 1,2; \; k = 1,2,3.$$

In words, the probability of a randomly selected observation being in any cell ijk is the product of the three probabilities. The expected cell count under this model would be

$$m_{ijk} = Np_{ijk} = Np_{i..}p_{.j.}p_{..k},$$

which illustrates the multiplicative relationship mentioned previously. An alternative formulation of m_{ijk} is

$$m_{ijk} = N \frac{m_{i++}}{N} \frac{m_{+j+}}{N} \frac{m_{++k}}{N} = \frac{m_{i++} \; m_{+j+} \; m_{++k}}{N^2}$$

where

$$m_{i++} = \sum_j \sum_k m_{ijk} = Np_{i..},$$

$$m_{+j+} = \sum_i \sum_k m_{ijk} = Np_{.j.}$$

and

$$m_{++k} = \sum_i \sum_j m_{ijk} = Np_{..k}.$$

Taking logarithms and letting ℓ_{ijk} represent the log of the expected cell count gives the loglinear model,

$$\ell_{ijk} = \ln(m_{ijk}) = \ln \frac{m_{i++}\, m_{+j+}\, m_{++k}}{N^2}$$

$$= -\ln(N^2) + \ln(m_{i++}) + \ln(m_{+j+}) + \ln(m_{++k}).$$

Recall that this model assumes independence of all three categorical variables. The four terms in the model can be thought of as representing an overall effect and main effects of the three categorical variables and is usually written as

$$\ell_{ijk} = \mu + \mu_{T(i)} + \mu_{R(j)} + \mu_{A(k)}, \qquad (1)$$

where $\mu = \frac{1}{12}\ln(m_{+++})$, $\mu_{T(i)} = \frac{1}{2}\ln(m_{i++}) - \mu$, $\mu_{R(j)} = \frac{1}{2}\ln(m_{+j+}) - \mu$ and $\mu_{A(k)} = \frac{1}{3}\ln(m_{++k}) - \mu$. As in the analysis of variance, these parameters are not estimable without restrictions which are

$$\sum_i \mu_{T(i)} = \sum_j \mu_{R(j)} = \sum_k \mu_{A(k)} = 0.$$

If the assumption of independence of all three variables is removed, additional terms are added to the model in order to account for these dependencies. All LL models are special cases of a general loglinear model (Fienberg, 1977) written as

$$\ell_{ijk} = \mu + \mu_{T(i)} + \mu_{R(j)} + \mu_{A(k)} + \mu_{TR(ij)} + \mu_{TA(ik)}$$

$$+ \mu_{RA(jk)} + \mu_{TRA(ijk)}. \qquad (2)$$

Model (2) is often referred to as the saturated model because it contains exactly as many parameters to be estimated as there are cells in the table [i.e., [1 + (I - 1) + (J - 1) + (K - 1) + (I -1) (J - 1) + (I - 1) (K - 1) + (J - 1) (K - 1) + (I - 1) (J - 1) (K - 1)] versus [IJK]}; hence, there are no degrees of freedom in (2) for tests of lack of fit. (Degrees of freedom = # cells - # parameters estimated).

By setting various terms in model (2) equal to zero, other models which fall in between the independent case of model (1) and the saturated case of model (2) are possible. For example, a simpler model than model (2) can be postulated by setting $\mu_{TRA} = 0$ and calculating a goodness-of-fit statistic for the expected cell counts from the hypothesized model. If this statistic is significant at some given level, the conclusion is that the simpler model does not fit the data observed. The dependencies must involve all three variables in a significant, triple interaction term.

We denote the likelihood-ratio chi-square statistic for the test of $\mu_{TRA} = 0$ by G^2_{TRA}. If this statistic is nonsignificant, usual practice is to fit a model without the three-way interaction but including any two-way interaction terms of interest. In this way, many other models are possible as long as two "rules" are obeyed. First, the formulation of alternative models is limited to those included in "hierarchical sets in which higher-order terms may be included only if the related lower-order terms are included" (Fienberg, 1977, p. 39). An example of two models that would be included in the same hierarchical set would be models (3) and (4) below because

$$\ell_{ijk} = \mu + \mu_T + \mu_R + \mu_A + \mu_{TA} + \mu_{RA} \qquad (3)$$

and

$$\ell_{ijk} = \mu + \mu_T + \mu_R + \mu_A + \mu_{TA} + \mu_{RA} + \mu_{TR} \qquad (4)$$

follow this rule while, for example

$$\ell_{ijk} = \mu + \mu_T + \mu_R + \mu_{RA}$$

and

$$\ell_{ijk} = \mu + \mu_A + \mu_{TR}$$

do not.

The second requirement is that terms involving fixed marginal totals must be included in the model being tested. Within the context of the present problem, the marginals are fixed for test and ability; therefore, the terms μ_T, μ_A and μ_{TA} must be included.

As mentioned previously, a significance test of a postulated model is in the form of a goodness-of-fit test via a likelihood-ratio chi-square statistic, denoted by G^2. As each term is added to the model and the model is tested against the observed data, a G^2 statistic is obtained. The difference between the G^2 statistic for this model and the previous model is then a chi-square statistic for a test of the significance of the added term. This procedure of successively fitting hierarchical models is more completely discussed by Bishop, Fienberg and Holland (1975) and Fienberg (1977). These sources also give methods for computing expected cell counts for each hypothesized model, as well as the constraints imposed.

For the present problem, there are three model possibilities to consider. The first case is given by model (3) and implies that the only terms needed to describe both pretest and national tests, other than the fixed marginal terms that must be added, are μ_R and μ_{RA}. The former term is a function of the item's overall difficulty level while the response by ability interaction models the item's discrimination characteristics. Certainly we want $\mu_{RA} \neq 0$; otherwise, the item has no overall discrimination at all.

If no other terms are necessary, then model (3) describes "no change" in either difficulty or discrimination from pretest to national administrations (i.e., no terms involving T have to be added other than those required). However, if there are changes in overall difficulty between administrations, then the μ_{TR} will have to be added, giving model (4). The difference in G^2 statistics between (3) and (4) or G^2_{TR} tests the significance of the added term with degrees of freedom (d.f.) equal to the difference in d.f. of the two models, or $(I - 1)(J - 1) = 1$ in this case.

By adding one more term to model (4), we get the saturated model or model (2). If $\mu_{TRA} \neq 0$, then there are differences in discrimination between test administrations. Because the saturated model (2) has zero d.f., the significance of μ_{TRA} is basically equivalent to the situation described previously; namely, if $\mu_{TR} \neq 0$ and the model still doesn't fit the data, (i.e., model (4) is significant) then a three-way interaction must be added to describe the observed cell counts, and $\mu_{TRA} \neq 0$. So the test of model (4) is a test of the significance of μ_{TRA}. It is possible to have a model that describes discrimination differences but no overall differences in p. If model (3) did not fit the data and if model (4) did not fit the data, but G^2_{TR} were not significant, then μ_{TRA} must be significant. In this situation, there would be item discrimination differences between testing administrations, but there would be no overall item difficulty differences. Or,

there could exist overall item difficulty and discrimination differences between pretest and national samples. The hierarchical principle of LL model testing does not require that $\mu_{TR} \neq 0$ before μ_{TRA} can be tested.

Logistic Regression

If the expected values of the cell counts of the two levels of response, m_{i1k} and m_{i2k}, were written as a ratio, m_{i1k}/m_{i2k}, where level 1 of response was correct and level 2, incorrect, then an equivalent way to write the LL model (2) would be as

$$\ln\left(\frac{m_{i1k}}{m_{i2k}}\right) = \{\mu - \mu\} + \{\mu_{T(i)} - \mu_{T(i)}\} + \{\mu_{R(1)} - \mu_{R(2)}\} + \{\mu_{A(k)} - \mu_{A(k)}\}$$

$$+ \{\mu_{TR(i1)} - \mu_{TR(i2)}\} + \{\mu_{TA(ik)} - \mu_{TA(ik)}\} + \{\mu_{RA(1k)} - \mu_{RA(2k)}\}$$

$$+ \{\mu_{TRA(i1k)} - \mu_{TRA(i2k)}\}$$

$$= \{\mu_{R(1)} - \mu_{R(2)}\} + \{\mu_{TR(i1)} - \mu_{TR(i2)}\} + \{\mu_{RA(1k)} - \mu_{RA(2k)}\} +$$

$$\{\mu_{TRA(i1k)} - \mu_{TRA(i2k)}\},$$

or

$$\ln\left\{\frac{m_{i1k}}{m_{i2k}}\right\} = w + w_{T(i)} + w_{A(k)} + w_{TA(ik)}$$

where

$$w = \{\mu_{R(1)} - \mu_{R(2)}\},$$

$$w_{T(i)} = \{\mu_{TR(i1)} - \mu_{TR(i2)}\},$$

$$w_{A(k)} = \{\mu_{RA(1k)} - \mu_{RA(2k)}\},$$

and

$$w_{TA(ik)} = \{\mu_{TRA(i1k)} - \mu_{TRA(i2k)}\}.$$

The logarithm of the ratio, m_{i1k}/m_{i2k}, is called a logit. Fienberg (1977) has described logit models as the categorical response analogs to regression models for continuous variables. If all the variables in a LL analysis are fixed except for a single response variable, then logit models and loglinear models lead to the same, exact results. Item bias detection studies using logit models rather than LL models have been described by Kok et al. (1985), Van Der Flier et al. (1984) and Mellenbergh (1982). Logit models are easiest to interpret when the response variable is dichotomous.

If at least one of the fixed or explanatory variables is continuous, the contingency table approach will not work unless the continuous variable is converted to a categorical variable. Rather than force the loss of information and statistical power by arbitrarily categorizing a continuous variable, we can write the logit model as

$$\ln \left\{\frac{p_i}{1 - p_i}\right\} = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}, \; i = 1,2,\ldots, N$$

where

X_{i1} = ability score for the ith examinee,

X_{i2} = test administration variable, a categorical variable where

$$X_{i2} = \begin{cases} -1 \text{ if examinee is in the national test sample,} \\ +1 \text{ if examinee is in the pretest sample,} \end{cases}$$

X_{i3} = ability by test interaction, $X_{i1} \bullet X_{i2}$,

and

p_i = the ith examinee's probability of responding correctly, given X_{i1}, X_{i2}, and X_{i3}.

This model is linear in the logit metric but nonlinear or logistic when written as

$$p_i = \frac{\exp\{\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}\}}{1 + \exp\{\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}\}}. \qquad (5)$$

Model (5) is referred to as a logistic regression (LR) model, and estimates of the regression coefficients can be obtained through the method of maximum likelihood estimation (MLE) from the iterative solution of a system of nonlinear equations.

The significance test procedures for the regression coefficients parallel those for the LL terms. If the observed response data fit a LR model where $\beta_2 = \beta_3 = 0$, then the resulting model is analogous to LL model (3) in that a single logistic curve,

$$p_i = \frac{\exp\{\beta_o + \beta_1 X_{i1}\}}{1 + \exp\{\beta_o + \beta_1 X_{i1}\}} \qquad (6)$$

describes the probability of a correct response, given only an ability score, X_{i1}. No differences would be observed between pretest and national administrations. If there were overall item difficulty differences between the two tests, the β_2 term would be significantly different from zero and

$$p_i = \frac{\exp\{\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2}\}}{1 + \exp\{\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2}\}}. \qquad (7)$$

The addition of the β_2 term changes the constant term, β_0 for each group but leaves the "slope" or steepness coefficient equal to β_1 for both groups (i.e., overall item difficulty has changed but discrimination has not). If β_3 were also significant, the ability by test interaction would change for each test, implying that there were significant changes in item discrimination. This would yield LR model (5), analogous to the saturated LL model (2).

Significance tests of the LR model are again carried out with the use of likelihood-ratio chi-square statistics, G^2, for goodness-of-fit tests. Differences in G^2 values of hierarchical models test the significance of the added term. In the LR situation, however, the degrees of freedom of the model tests are equal to the number of patterns observed (i.e., "cells" or combinations of the observed independent variables) minus the number of parameters or coefficients estimated.

Confidence bands around the logistic response surface can be constructed according to a method proposed by Hauck (1983). This method produces $100(1 - \alpha)\%$ confidence statements that apply to all X_{ij}, $i = 1,\ldots, N$; $j = 1, 2, 3$ in LR model (5). Therefore, the goodness-of-fit of the model, overall, can be tested with the likelihood-ratio chi-square statistic, while comparisons of the confidence bands for p_i for each test administration at each ability level can be made in a post hoc manner, similar to a Scheffé procedure.

Method

These two procedures, LL and LR, were compared on their ability to detect item difficulty and discrimination changes of 40 ACT Assessment Mathematics Usage items between the time that each item was pretested until the items appeared in a final form as part of a national test administration. These 40 items had been pretested in 1982 on previous ACT Assessment administrations to a total of 7127 examinees. Because mathematics items are pretested in multi-item units, it was possible that some of the examinees responded to more than one of the items that eventually comprised the 40 items selected for a national form.

A random sample of 8000 was selected from the 131,000 examinees who took a 1984 national form of the ACT Assessment, and this became the national test administration sample in this comparison study.

Because items would be analyzed separately (i.e., there would be 40 separate LL and LR analyses to perform), we divided the national sample of 8000 into 40 subsamples of size 200 each. This was done in order to (a) make sure that the 40 tests or analyses would be independent and (b) to make the national samples similar in size to the pretest samples since pretest samples for these 40 items ranged from 179 to 248 (ave. = 210). Pretest sample units in which examinees responded to more than one item were subdivided to insure that, similar to the national sample, all 40 pretest samples were independent. This meant that a few of the pretest samples were halved or even divided into thirds to insure this independence.

Table 1 gives the item difficulty and discrimination parameters in terms of p and the biserial correlation coefficient, R. The latter is computed between the ACT Assessment Mathematics Usage equated scores and the test item response. Mean p and R values for both test administrations are listed at the bottom of columns 3-6. In general, according to these parameter estimates, the items on

this 40-item test tended to become slightly easier and more discriminating on the final version of the test. However, these are fairly general observations based on item statistics that may or may not be sensitive to real item characteristic changes. The last column in Table 1 lists whether or not an item was modified between test administrations. Usually math items receive very slight modifications after pretesting if they receive any at all.

Because the examinees at each of the two administrations responded to the same item but took different forms of the ACT Assessment, the Mathematics Usage equated score was used as an examinee's ability score. This equated or standard score ranges from 1 to 36. For the LL analyses, the 40 items were first analyzed using three ability levels, as pictured in Figure 1, in the BMDP4F computer program for multi-way contingency tables (BMDP, 1983a). To form the ability categories, the standard scores of the 8000 examinees from the national administration were put into a frequency distribution. The score at approximately the 33-1/3 percentile rank was used as the first cutoff score ($X_1 \leq 13$), scores between the 33-1/3 and 66-2/3 percentile ranks ($22 \geq X_1 > 13$) were used to classify examinees into the Medium ability category; and scores above the 66-2/3 percentile rank ($X_1 > 22$) defined the High ability category. The LL analysis with three ability categories is abbreviated as LL3.

Two other categorizations and analyses, one with four ability categories (LL4) and one with five (LL5) were also performed. The percentile rank (PR) scores and corresponding cutoff scores for LL4 were the 25th PR, the 50th PR, and the 75th PR corresponding to the standard scores, 11, 19, and 24, respectively. For the LL5 analysis, the 20th PR, 40th PR, 60th PR and 80th PR corresponding to standard scores 10, 16, 21, and 25, were used.

As mentioned previously, tests of item difficulty changes and/or item discrimination changes were made via tests on the significance of the difference between the G^2 statistics of various models. To test item difficulty changes, we

tested the significance of $G^2_{(3)} - G^2_{(4)}$ or G^2_{TR} with one d.f. To test item discrimination changes, we tested $G^2_{(4)}$ or G^2_{TRA} with (k - 1) d.f., where k = 3,4, or 5 ability categories.

In the logistic regression (LR) analysis, the equated math score itself was used as an examinee's ability score, X_{i1}. Model (5), the fullest LR model, was first fit to the item responses of both groups combined, using the BMDPLR stepwise logistic regression program (BMDP, 1983b) with independent variables X_{i1}, X_{i2}, and X_{i3}, described previously. In a forced, stepwise procedure, the removal of the X_{i3} and X_{i2} variables from the model tested the significance of β_3 and β_2 via the differences in G^2 statistics of models (5) and (7). Therefore, $G^2_{(6)} - G^2_{(7)}$ or G^2_T gave a test of item difficulty changes while $G^2_{(7)} - G^2_{(5)}$ or G^2_{TA} gave a test of item discrimination changes. Both tests had one degree of freedom.

Results and Discussion

Table 2 gives the results of the LL and LR analyses for item difficulty and item discrimination changes. These results are in the form of the G^2 model difference or model improvement statistics (hereafter referred to as the G^2-improvement statistic). The statistics for Table 2 for item difficulty (columns 2-5) are all distributed as asymptotic chi-square random variables with one degree of freedom. The G^2-improvement statistics for discrimination (columns 6-9) are all distributed as asymptotic chi-square random variables with either one, four, three or two d.f., for LR, LL5, LL4 and LL3 procedures respectively.

It can be seen that only one test item, #28, was significantly different on overall item difficulty from pretest to national administrations. The item was identified by the LR and by LL5, LL4 and LL3 procedures as well. The significance level for all tests was taken to be $p < .0005$, corresponding approximately to an overall Type I error rate of .05 for 80 tests (i.e., 40 test items with tests of difficulty and discrimination changes per item).

Similarly, only one test item, #35, showed a significant change in item discrimination from pretest to national administrations. Again all of the procedures "flagged" this test item.

To observe how consistent these findings were for all of the items, the G^2-improvement statistics for difficulty were ranked within each method and the ranks were correlated. Similar rankings and correlations were performed on the discrimination statistics. The resulting rank correlation matrices appear in Table 3.

The size of the correlation coefficients supported the intuitive notion that, as long as ability was not an issue (i.e., for tests of overall difficulty), all of the methods would yield approximately the same results. However, as soon as the ability variable became a factor, the procedures were no longer as consistent. Those procedures which forced the categorization of the "continuous" ability variable did not produce results that were strongly consistent with the LR procedure, one that did allow ability to be treated as continuous. Even the three LL procedures were not in strong agreement among themselves, except for LL4 and LL5.

The LR procedure leads to a graphical display of the item changes since LR model (5) can be plotted once estimates of β_0, β_1, β_2, and β_3 are known. For each test administration, substitution of the appropriate value (± 1) for X_{i2} and X_{i3} yielded estimated item characteristic functions. Examples of these plots are shown in Figures 2 and 3 for Item #28 and #35, respectively.

It is interesting to note that although these procedures did identify that two items had significantly "changed" item characteristics between test administrations, very little, if any, modifications were done to the items after pretesting. Item #28 wasn't modified at all (see Table 1, column 7) and #35 received very slight wording changes.

The list of alternatives was unchanged. However on the pretest unit, the item appeared towards the beginning while on the national form, it was towards the end of the test. But these were the only noticeable differences between the two items. And yet the item's discrimination improved drastically, as reinforced by Figure 3.

Conclusions

Both the LL and LR procedures consistently located those test items which had apparently changed in item difficulty (#28) and item discrimination (#35) from pretest to national test administrations. We say "apparently" because this was not a simulation study and we do not know with certainty that these two items actually did change in terms of these characteristics. From a production cost standpoint, the performance of the LL analyses is encouraging since these tests, in terms of computer costs, are less expensive than the LR procedure. Each LL analysis for all 40 items was approximately $7.50 in terms of total run costs (CPU = 13.5 sec.) while the 40 LR analyses, in total, cost $30.00 (CPU = 55.5 sec.).

Aside from cost, another advantage of the LL procedure, is that the LL models do not impose a logistic, monotonic model assumption on the responses as a function of ability. These models would therefore appear to handle nonmonotonic situations better than LR models. On the other hand, when tests of item discrimination differences are important, the treatment of the ability variable as a continuous variable measured on an interval scale rather than as a nominal variable might outweigh the computer cost differences, if the data did fit the logistic function well. And the use of the confidence intervals around each LR curve might prove to be useful at particular ability scores of interest. Finally, the LR procedure certainly has some similarities to an IRT (item response theory) approach in terms of the resulting item characteristic

curves (ICC's) that are estimated. Some preliminary analyses have indicated that, on ACT Assessment Program test data, an LR analysis produces estimates of item difficulty ($-\hat{\beta}_0/\hat{\beta}_1$) and item discrimination ($\hat{\beta}_1$) that correlate moderate-to-high with **b** and **a** parameters from a 2-parameter logistic model (.95 and .59, respectively). The exact duplication of parameter estimation by the LR procedure may not be as important, however, as the detection of significant ICC differences or changes. Further studies need to be conducted to see if the LR and LL procedures are sensitive to such significant changes in ICCs from a latent trait or ability scale. Our preliminary results on real test data, however, were encouraging.

REFERENCES

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

BMDP (1983a). Frequency tables [Computer statistical software] Berkeley, CA: University of California Press (BMDP4F).

BMDP (1983b). Stepwise logistic regression [Computer statistical software] Berkeley, CA: University of California Press (BMDPLR).

Fienberg, S. E. (1977). The analysis of cross-classified categorical data. Cambridge, MA: MIT Press.

Hauck, W. W. (1983). A note on confidence bands for the logistic curve. The American Statistician, 37, 158-160.

Kelderman, H. (June, 1985). Item bias detection using the loglinear Rasch model: observed and unobserved subgroups. Paper presented at the annual meeting of the Psychometric Society, Nashville, TN.

Kok, F. G., Mellenbergh, G. J., & Van Der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22, 295-303.

Marascuilo, L. A. & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. Journal of Educational Measurement, 18, 229-248.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Rentz, R. R. (1978, March). Monitoring the quality of an item pool calibrated by the Rasch model. Paper presented at the meeting of the National Council on Measurement in Education, Toronto, Ont.

Searle, S. R. (1971). Linear models. New York: Wiley.

Van Der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

TABLE 1

Classical Item Difficulty and Discrimination Estimates at Pretest and National Test Administrations

Item #	Pretest Sample Size	Pretest p	Pretest R	National p	National R	Modified ?
1	116	.78	.53	.78	.62	No
2	113	.75	.65	.79	.60	No
3	105	.71	.54	.73	.58	Yes
4	238	.66	.45	.64	.55	No
5	90	.65	.68	.66	.65	Yes
6	227	.61	.41	.70	.55	Yes
7	216	.60	.36	.72	.53	No
8	71	.59	.59	.58	.68	Yes
9	219	.57	.58	.54	.64	Yes
10	99	.57	.57	.55	.49	Yes
11	71	.57	.43	.63	.47	Yes
12	233	.55	.46	.69	.63	Yes
13	116	.53	.57	.53	.58	No
14	195	.53	.34	.52	.50	Yes
15	231	.53	.34	.63	.48	No
16	198	.50	.46	.60	.56	Yes
17	206	.50	.37	.53	.58	Yes
18	205	.47	.53	.55	.64	Yes
19	202	.46	.59	.55	.66	Yes
20	104	.45	.62	.55	.57	No
21	70	.45	.41	.51	.57	No
22	192	.45	.39	.47	.43	No
23	89	.44	.68	.52	.67	Yes
24	199	.43	.34	.46	.52	No
25	194	.42	.37	.39	.57	Yes
26	231	.41	.41	.45	.63	Yes
27	244	.40	.38	.52	.40	Yes
28	229	.39	.46	.55	.68	No
29	214	.37	.61	.47	.70	Yes
30	225	.37	.53	.48	.66	Yes
31	248	.37	.41	.37	.52	No
32	113	.35	.51	.31	.76	Yes
33	210	.35	.34	.43	.51	No
34	229	.34	.50	.41	.57	No
35	215	.34	.32	.37	.58	Yes
36	229	.32	.44	.39	.59	No
37	232	.31	.37	.30	.55	Yes
38	227	.30	.47	.41	.64	No
39	98	.30	.47	.39	.63	Yes
40	184	.29	.69	.29	.56	No
Total	7127					
Mean		.47	.48	.52	.58	

TABLE 2

Likelihood-ratio Chi-square Statistics for Logistic Regression and Loglinear Analyses

	DIFFICULTY				DISCRIMINATION			
Item	LR	LL5	LL4	LL3	LR	LL5	LL4	LL3
1	0.371	0.350	0.480	0.40	0.611	2.08	0.52	1.89
2	0.371	0.050	0.030	0.02	0.911	3.57	1.80	0.92
3	0.133	0.100	0.001	0.18	0.169	2.12	1.83	2.03
4	0.384	0.300	0.530	0.27	3.726	10.36	4.38	1.12
5	1.803	2.200	1.420	3.05	0.079	8.53	7.40	1.96
6	3.108	2.780	2.820	3.36	6.208	8.79	6.21	8.85
7	8.292	8.290	8.990	9.08	0.354	7.72	9.68	1.25
8	3.045	2.480	3.760	3.66	1.918	4.60	2.45	2.40
9	0.074	0.130	0.200	0.25	0.283	3.49	5.64	1.17
10	2.648	2.320	2.600	1.95	2.481	6.24	2.45	0.91
11	0.314	0.390	0.350	0.25	0.270	6.29	3.64	4.82
12	2.760	2.760	2.450	3.55	1.095	2.65	2.58	0.67
13	0.047	0.001	0.040	0.01	1.383	1.91	0.92	3.31
14	0.659	0.670	0.690	1.06	3.597	7.21	3.20	8.94
15	1.905	2.530	1.780	1.62	1.443	2.66	2.19	2.81
16	6.088	5.780	4.760	4.30	2.872	8.97	9.35	9.69
17	0.500	1.160	0.340	0.60	2.525	8.63	5.34	1.79
18	0.210	0.500	0.050	0.16	0.259	3.07	0.92	0.45
19	1.107	0.590	0.910	1.37	0.464	6.01	7.20	1.58
20	0.015	0.010	0.010	0.21	0.107	1.86	2.85	0.10
21	1.995	1.380	1.820	1.72	1.760	1.96	1.48	0.72
22	1.623	2.360	2.340	1.40	0.231	2.04	0.27	0.44
23	0.057	0.070	0.380	0.14	1.545	4.79	2.81	1.99
24	1.292	1.860	1.640	1.36	1.233	5.22	5.58	1.48
25	3.792	5.530	5.020	4.36	0.202	0.19	0.77	1.46
26	0.139	0.290	0.690	0.20	0.260	8.17	6.29	9.34
27	2.531	2.250	2.140	4.40	0.358	3.78	1.93	4.15
28	**13.229**	**12.420**	**11.660**	**12.13**	1.139	3.37	1.21	1.20
29	2.447	4.880	3.010	5.21	7.184	6.25	6.34	6.22
30	1.785	0.590	1.550	1.75	7.354	4.98	5.27	6.78
31	0.932	0.970	0.740	0.74	2.459	9.23	6.06	5.45
32	2.985	4.410	5.310	1.59	4.527	2.65	3.26	2.60
33	7.061	5.580	5.810	7.54	3.538	4.39	3.00	1.86
34	0.048	0.070	0.010	0.23	2.569	8.92	4.71	0.23
35	0.025	0.001	0.110	0.04	**19.550**	**16.42**	**12.17**	**9.25**
36	0.783	0.110	0.940	0.19	6.475	3.84	2.46	2.88
37	0.530	0.100	0.150	0.47	0.730	3.42	0.64	0.21
38	0.083	0.140	0.030	0.08	2.310	6.74	2.62	5.90
39	0.004	0.020	0.100	0.02	1.490	4.43	1.05	1.28
40	3.648	4.130	3.250	1.35	0.022	5.13	1.85	1.07
(D.F.)	(1)	(1)	(1)	(1)	(1)	(4)	(3)	(2)

TABLE 3

Rank Correlation Matrices of the G^2-improvement Statistics of the LR and LL Procedures

Difficulty

	LR	LL5	LL4	LL3
LR	1.000			
LL5	.946	1.000		
LL4	.945	.927	1.000	
LL3	.916	.913	.889	1.000

Discrimination

	LR	LL5	LL4	LL3
LR	1.000			
LL5	.435	1.000		
LL4	.343	.766	1.000	
LL3	.441	.429	.461	1.000

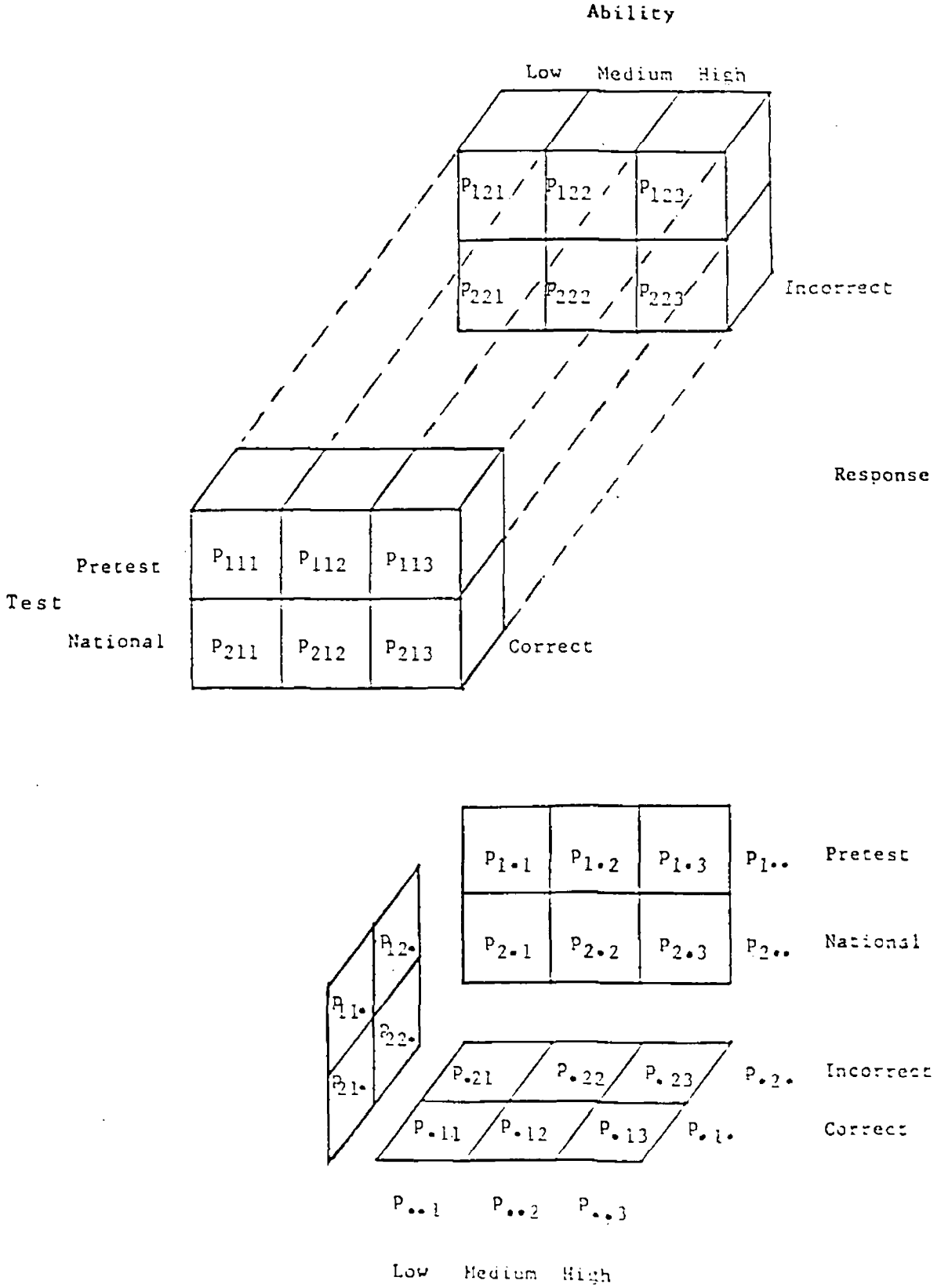

Figure 1. Proportions of Examinees By Ability, Response and Test

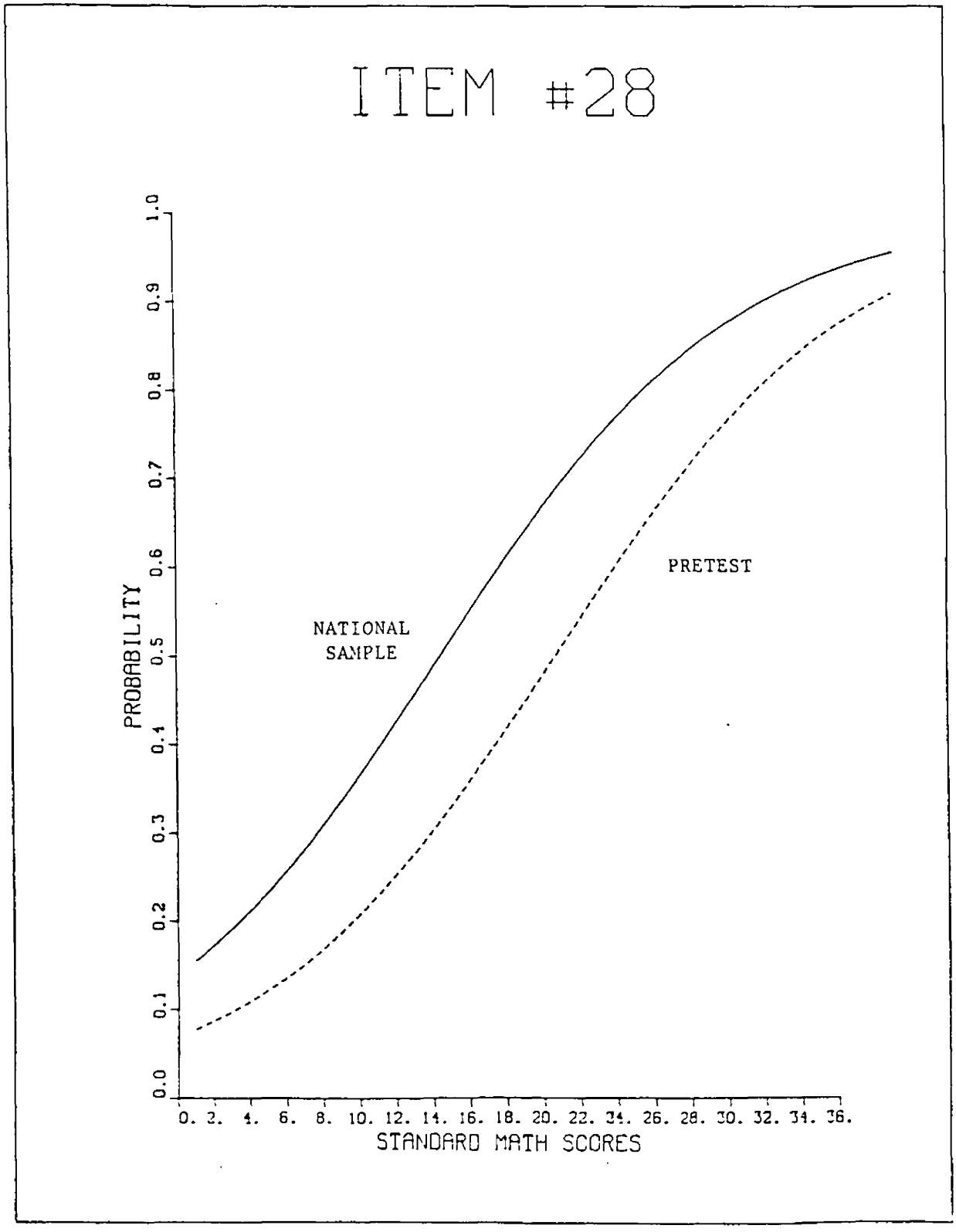

Figure 2. LR Curves of Item #28

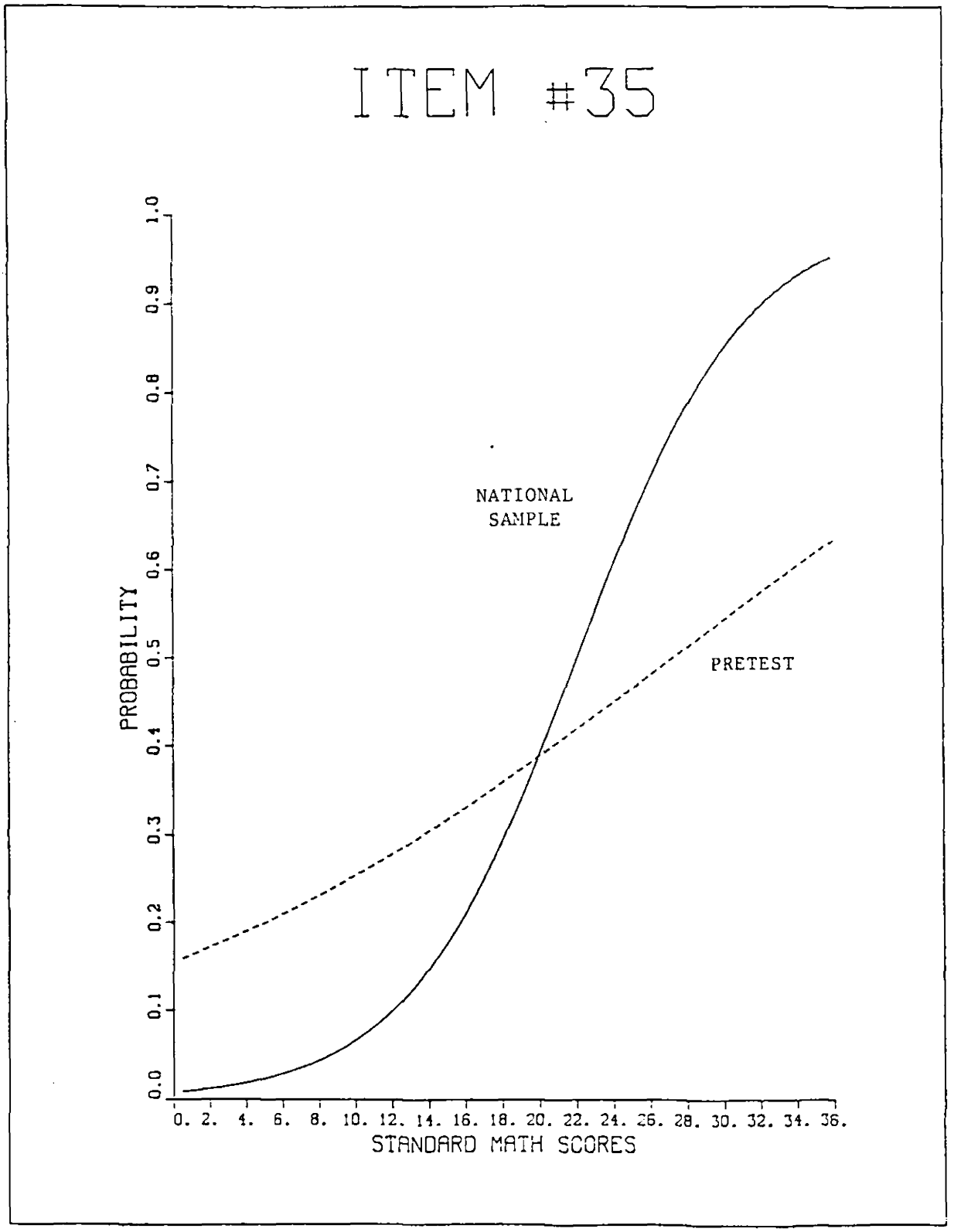

Figure 3. LR Curves of Item #35