MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 15, NUMBER 2, WINTER 1987

Using Diagnostics for Identification of Biased Test Items

Donald T. Searls
University of Northern Colorado
Edgar Ortiz
Citicorp

ABSTRACT

This paper demonstrates how recent developments in the analysis of regression models may prove useful in the identification of atypical and potentially biased test items. Regression diagnostics studied are based on analysis of the sensitivity of leverage points, studentized residuals, and ratios of covariances due to the sequential deletion of each test item from the analysis. These procedures appear to offer a substantial refinement over existing approaches.

IDENTIFICATION OF INFLUENTIAL ITEMS : THEORETICAL RATIONALE

Many statistical procedures have been proposed for detecting biased items. Although they differ in their conceptualization of bias, they nevertheless exhibit a commonality in their purpuse which is to identify those items which hamper the performance of one group relative to another.

Irrespective of the approach, the proposed statistical procedures for identifying biased items rely directly or indirectly on variants of the concept of statistical distance. A major limitation with all of these approaches is that no distribution theory is available to determine objectively when one atypical score is statistically different from others. This shortcoming is particularly evident in Angoff's delta-plot method and extensions of this procedure (Angoff and Ford, 1973. Rudner, et al., 1980).

A lack of distribution theory is also evident in the chi-square methods of Scheuneman (1979) and Camilli (1979). These procedures aim at detecting biased items by performing tests of randomness on the distribution of responses into ability intervals. However, setting of cut-off levels to establish the various ability intervals is done after examining the data. Such a posteriori determination of cutoff points to define ability intervals in effect violates the assumption of random assignment, since factors other than chance are influencing the results. Consequently, rather than detecting biased items, results so derived may identify instead an item's sensitivity to clustering into the ex post facto determined ability classes.

Statistical procedures for detecting biased items based on latent trait models have also been proposed. (Lord and Novick, 1968; Hambleton and Cook, 1977). In these methods, item characteristic curves are fitted to the observed performance scores of different groups. If the fitted curves are not the same for the groups being compared, the item is said to be biased. A major shortcoming of this approach is the lack of specification of the underlying theoretical distributon of the observed delta-values that characterize the differences in performance between the groups being compared. Although some progress has been reported (Lord, 1977), the validity of tests of significance to identify biased items based on the assumptions of latent trait models is as yet an issue that remains unresolved (Lord, 1977; p. 25). A comparative analysis of the performance of latent trait models to identify biased items (Rudner, et al. 1980), does not deal with the subject of statistical significance of the various indices of bias reported in that study.

A comprehensive review of the various statistical techniques proposed for detecting item bias is given in Peterson (1977), Merz (1978) and Sheppard et al. (1980). Statistical analyses, however, do not detect biased items. They only identify those items in which the achievement scores of the groups being compared deviate from the pattern established by other items that make up a test. These items, in turn, may reveal specific content characteristics that either increase or decrease the a priori probability of a correct response in one group of examinees but not in the other.

The statistical procedures to be exemplified in this investigation offer an objective set of statistical criteria to examine individual items for potential bias. These methods are based on generalizations of regression models as developed by Belsley, Kuh and Welsch (1980). The identification of potentially biased items, based on regression diagnostics offers a substantial refinement over existing approaches in that :

a) Distribution theory is used to determine cutoff levels and identify atypical items objectively.

b) Statistical methods are available that measure the sensitivity of parameter estimates to perturbartions in the data, e.g. the effects of the deletion of each item on the estimates of the regression coefficients.

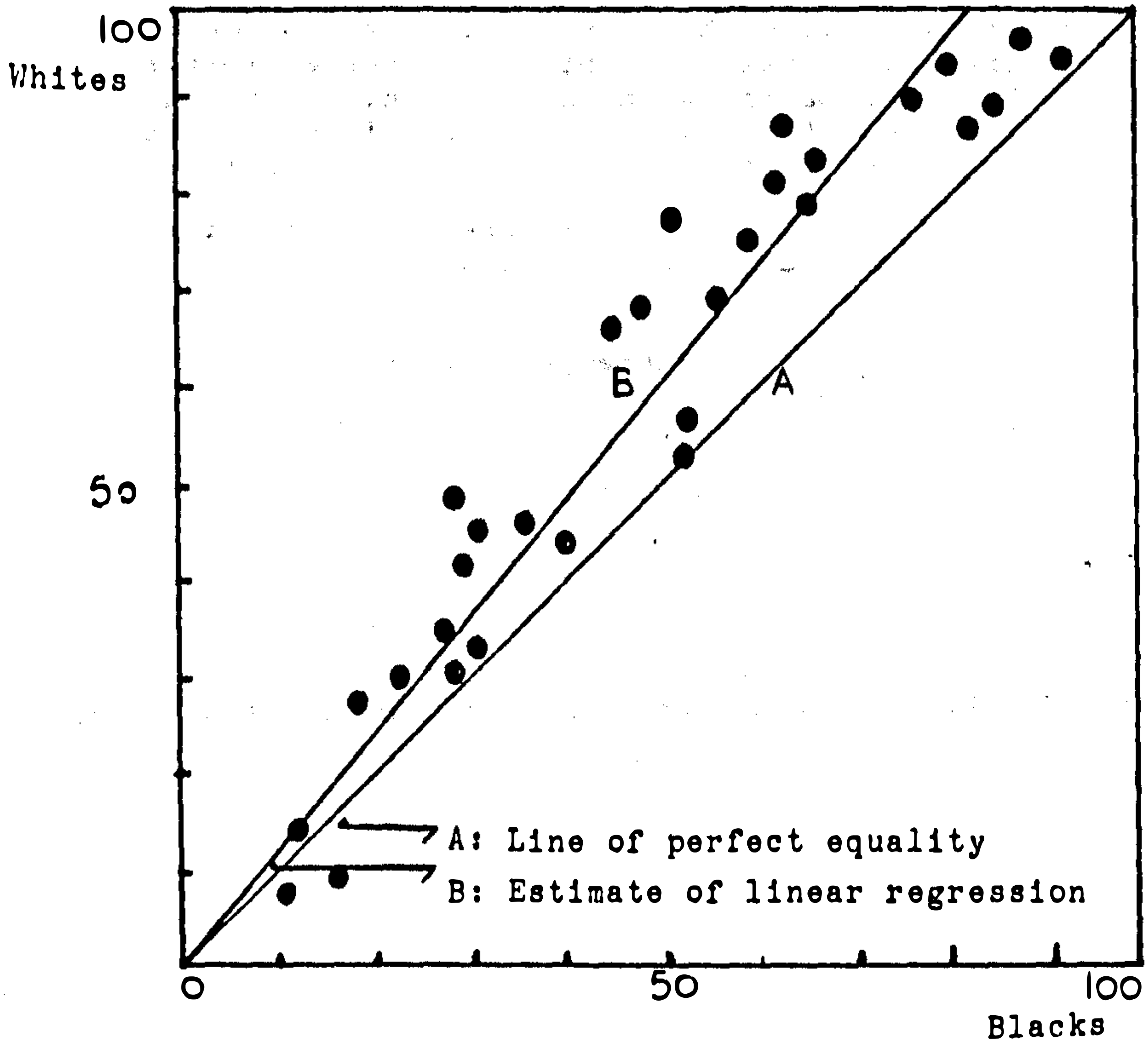
c) These methods offer measures of statistical distance independent of sample size.

Analysis of data based on these procedures can yield important information concerning atypical items which cannot be readily obtained by means of delta-plot, chi-square and latent trait models.

The data to be analyzed comprise the proportion of white and black students who attempted and responded correctly (p-values) to an assessment booklet consisting of 30 items. A scatter plot of the p-values is given in figure 1. Points on line A correspond to items in which the performance of both groups was equal. Points lying above and below this line correspond to items in which the groups being compared performed differently. Points above this line correspond to items in which the group represented by the vertical axis, performed better than the group represented by the horizontal axis. Similarly, points lying below this line correspond to items in which the group represented by the horizontal

FIGURE 1

PLOT OF ACHIEVEMENT SCORES OF
WHITE AND BLACK EXAMINEES



axis performed better than the group represented by the vertical axis.

An estimate of the regression line is given by line B (slope=1.19, p=.0001). From graph 1, the consistent scatter of points above line A indicates that white examinees have performed consistently above the performance level set by black examinees. The dispersion pattern of p-values around this line suggests a strong curvature at both extrema, i.e., in the range of the easiest and most difficult exercises. In order to correct for these bottom and ceiling effects, the the p-values were transformed to logits. The logistic transformation is widely used in the analysis of proportional data. Reexpressing quantal response data in logits provides a straightforward procedure to correct for interaction often found in exercise data in the easy and difficult range.

The techniques to be exemplified in this investigation, aim at identifying potentially biased items, by measuring the sensitivity of regression models to the deletion of individual items from the bulk of the data. These diagnostic methods will be applied to parameter estimates in regression models relating the performance of white and black examinees with p-values transformed into logits. Items whose deletion from the body of the data, cause atypical perturbations on parameter estimates are suspect.

For example, given a simple bivariate regression model, the magnitude of the perturbation on the estimated regression coefficients due to deletion of the ith item, can identify atypical items which warrant further examination for potential bias. This procedure is akin to estimating N regression models, where each model corresponds to the 'not i observation'. Within the context of our investigation, items whose deletion cause large and atypical perturbations on estimates of the regression parameters are therefore suspect. From a practical viewpoint this procedure is equivalent to a pseudo-experiment in which it is asked, how would white and black examinees have performed if the ith item had been deleted from the assessment booklet? With these regression diagnostics, items having large deviations from the performance pattern observed in the remaining items can be readily identified.

RESULTS

DETECTION OF POTENTIALLY BIASED ITEMS BASED ON REGRESSION DIAGNOSTIC PROCEDURES

The regression diagnostics to be exemplified for use in the detection of potentially biased items are based on analysis of the sensitivity of leverage points, studentized residuals, and ratios of covariances due to the sequential deletion of each item from the model. Two regression models are examined. In model 1, the achievement scores of white examinees are predicted based on the performance of black examinees. Similarly, in model 2, the achievement scores of black examinees are predicted based on the performance of white examinees. The proposed diagnostics attempt to detect biased items by identifying those items that in either model 1 or model 2 elicit performance scores significantly different from the pattern of variability established in the remaining items that make up the achievement booklet. These diagnostic statistics follow from the usual linear model

$$Y = X\beta + e \qquad e \sim N(0,\sigma^2) \qquad (1)$$

where Y is a (n x 1) vector of observations on the dependent variable, X is a (n x p) matrix of observations on the explanatory variables, B is a (p x 1) vector of unknown regression parameters, and e is a (n x 1) vector of random errors. From (1), the least squares estimate of the vector of regression coefficients is

$$B = (X'X)^{-1}X'Y \qquad (2)$$

The least squares projection matrix, often called the hat matrix, is of fundamental importance in the identification of items that elicit atypical performance scores between the groups being compared. The hat matrix is defined as

$$H = X(X^1X)^{-1}X^1 \qquad (3)$$

The diagonal elements of H, denoted h , measure the influence or leverage of the response variable y on its corresponding fitted value.

Results derived by Belsley, et. al., (1980), and Hoaglin and Welsch (1978) provide a statistical criterion to set cutoff levels to identify observations whose pattern of influence is atypical. Their results indicate that values of h larger than 2*(p/n) need further examination due to their unusually large influence on the hat matrix, H. Observations that exceed this cutoff level are often termed 'leverage points' in the statistical literature.

Values of the diagonal elements of the H matrix are recorded in column 1 of tables 1 and 2 respectively. An examination of these values indicates that the cutoff level of .133 is exceeded by items 1 and 14 in model 1, and items 13 and 14 in model 2. The quantitative influence of these items on other aspects of the regression analysis is examined further in the following sections of this investigation.

A common practice in the item bias literature has been that of identifying as biased those items with large residual values in fitted linear models. This approach fails to take into account the fact that the variances of the residuals are not constant, but a function of the X matrix. Therefore, results so derived may lead to unwarranted conclusions concerning their potential bias. To avoid the problems associated with the non-constancy of the variances of the residuals, atypical items can be identified by scaling the residuals by their respective variances. For these purposes the residuals can be modified in ways that enhance our ability to detect those items which elicit the statistically most dissimilar performance. This transformation of the residuals is illustrated next. From (1) a least squares fit produces residuals given by

$$e = (I - X(X^1X)^{-1}X^1) \tag{6}$$

and mean square residuals

$$s = \frac{e^1e}{n-p} \tag{7}$$

The variance-covariance matrix of estimates of the residuals is

$$\mathrm{Var}(e) = \sigma^2(I-H) \tag{8}$$

where H is the least squares projection matrix defined in (3). Standardizing the residuals by estimating σ^2 by the residual mean square based on regression estimates without the ith observation yields the ratio of 'studentized residuals',

$$e(i) = \frac{e(i)}{s(i)\sqrt{1-h_i}} \tag{9}$$

These residuals are distributed as a t-distribution with n-p-1 degrees of freedom. Therefore, if the Gaussian assumption holds, the significance of any one of these studentized residuals can be readily assessed from tabulated values of the t-distribution with n-p-1 degrees of freedom.

Estimates of the studentized residuals are listed in column 3 of tables 1 and 2. The magnitude of the studentized residual for items 1 and 26 consistently exceeds the critical value of 1.70 (t, 27 df alpha= .05). In this particular

TABLE 1

White Regression Model

Model 1

Item No.	Hat Matrix	Raw Resid.	Stdzed. Resid.	Covar. Ratio	DFFITS	DFBETAS Const.	DFBETAS Slop
1	0.20*	-1.07	-3.24*	0.70*	-1.65*	-0.71*	-1.5
2	0.03	- .56	-1.43	0.96	-0.27	-0.26	-0.0
3	0.03	- .11	-0.29	1.11	-0.05	-0.05	0.0
4	0.09	- .47	-1.22	1.06	-0.40	-0.24	-0.3
5	0.03	- .05	-0.14	1.11	-0.02	-0.02	-0.0
6	0.04	0.10	0.26	1.11	0.05	0.05	0.0
7	0.04	0.25	0.62	1.09	0.14	0.11	-0.0
8	0.09	- .61	-1.63	0.98	-0.54	-0.29	0.4
9	0.04	.08	0.19	1.12	0.04	0.03	-0.0
10	0.05	0.43	1.10	1.03	0.25	0.20	-0.1
11	0.04	0.31	0.77	1.07	0.16	0.14	0.0
12	0.03	0.28	0.70	1.07	0.13	0.13	-0.0
13	0.12	0.26	0.68	1.19	0.26	0.14	0.2
14	0.14*	-0.44	-1.20	1.13	-0.49	-0.22	0.4
15	0.04	0.32	0.81	1.06	0.16	0.15	0.0
16	0.05	- .12	-0.31	1.12	-0.07	-0.05	0.0
17	0.03	0.29	0.73	1.06	0.13	0.13	-0.0
18	0.03	0.17	0.43	1.10	0.08	0.08	0.0
19	0.08	0.21	0.55	1.14	0.16	0.10	-0.1
20	0.03	- .35	-0.88	1.05	-0.16	-0.16	-0.0
21	0.12	- .18	-0.47	1.20	-0.17	-0.08	0.1
22	0.06	0.24	0.62	1.12	0.17	0.12	0.1
23	0.03	- .00	-0.00	1.11	-0.00	-0.00	0.0
24	0.06	.02	0.07	1.15	0.01	0.01	-0.0
25	0.03	0.59	1.52	0.94	0.28	0.28	0.0
26	0.04	0.75	1.97*	0.85	0.41	0.37*	0.1
27	0.10	-0.43	-1.12	1.09	-0.38	-0.22	-0.3
28	0.05	- .26	-0.67	1.09	-0.15	-0.12	0.0
29	0.08	0.48	1.24	1.04	0.37	0.24	0.2
30	0.05	- .19	-0.47	1.11	-0.11	-0.08	0.0

TABLE 2

Black Regression Model

Model 2

Item	Hat	Raw	Stdzed.	Covar.		DFBETAS	
No.	Matrix	Resid.	Resid.	Ratio	DFFITS	Const.	Slope
1	0.10	1.00	3.90*	0.49*	1.35*	0.31	1.12*
2	0.03	0.46	1.42	0.96	0.27	0.27	-0.06
3	0.04	.05	0.16	1.11	0.03	0.03	-0.01
4	0.06	0.49	1.56	0.96	0.42	0.17	0.29
5	0.03	.06	0.20	1.11	0.03	0.03	0.00
6	0.04	- .03	-0.11	1.12	-0.02	-0.01	-0.01
7	0.04	- .25	-0.78	1.07	-0.16	-0.16	0.07
8	0.13	0.36	1.17	1.12	0.47	0.36	-0.40
9	0.04	- .12	-0.36	1.11	-0.07	-0.07	0.03
10	0.03	- .41	-1.26	0.99	-0.25	-0.25	0.09
11	0.05	- .19	-0.58	1.10	-0.14	-0.07	-0.08
12	0.03	- .22	-0.68	1.07	-0.12	-0.11	-0.02
13	0.14	- .05	-0.17	1.25*	-0.07	-0.01	-0.06
14	0.17	0.19	0.62	1.26*	0.29	0.20	-0.26
15	0.04	- .21	-0.65	1.09	-0.14	-0.08	-0.08
16	0.05	.03	0.10	1.13	0.02	0.02	-0.01
17	0.03	- .24	-0.73	1.07	-0.12	-0.12	-0.01
18	0.04	- .10	-0.31	1.11	-0.06	-0.04	-0.02
19	0.06	- .28	-0.86	1.09	-0.23	-0.21	0.16
20	0.03	0.29	0.88	1.05	0.16	0.16	-0.02
21	0.12	.00	0.01	1.23*	0.00	0.00	-0.00
22	0.07	- .10	-0.31	1.15	-0.09	-0.03	-0.07
23	0.03	- .02	-0.06	1.11	-0.01	-0.01	0.00
24	0.06	- .11	-0.34	1.13	-0.09	-0.08	0.06
25	0.04	- .46	-1.44	0.96	-0.30	-0.20	-0.13
26	0.06	- .55	-1.76*	0.92	-0.46	-0.19	-0.33
27	0.07	0.47	1.49	0.99	0.42	0.15	0.31
28	0.05	0.15	0.45	1.12	0.11	0.10	-0.07
29	0.10	- .27	-0.85	1.14	-0.30	-0.06	-0.25
30	0.05	.08	0.25	1.13	0.06	0.06	-0.04

case there is substantial agreement between those items with relatively large residuals, and those with relatively large studentized residuals. The magnitude of the studentized residuals associated with items 1 and 26 indicate that the performance of white and black examinees in these two items is significantly different from the performance pattern established in other items. And as such, these items warrant further examination for potential bias. The studentized residuals e(i) offer a substantial improvement over the usual analysis of raw residuals, both because they have equal variances and because an underlying distribution theory exists to identify atypical values.

Another important group of diagnostic methods measure the impact of the deletion of the ith observation on the stability of several statistical ratios, and estimated regression coefficients. Statistical procedures that have been developed to estimate the impact of the deletion of the ith observation on these statistics, are examined next. An important diagnostic statistic is the covariance ratio. This ratio is formed by comparing the covariance of the regression model whith the ith observation deleted, and the covariance of the complete regression model. By repeating this procedure for each observation in the sample, a set of N values that corresponds to estimates of the covariance ratios is obtained. Atypical items can be identified by measuring the impact of their deletion on the estimates of the covariance ratios. Covariance ratios based on the 'not ith' observation which deviate from one, indicate that this particular observation is exerting an atypical influence, and needs therefore further examination. From (1) the variance-covariance matrix of the regression coefficients is:

$$\mathrm{Var}(b) = \sigma^2 (X^1X)^{-1} \qquad (11)$$

Similarly, the variance-covariance matrix of the regression coefficients due to the 'not ith' observation is,

$$\mathrm{Var}(b(i)) = \sigma^2 (X^1(i)X(i))^{-1} \qquad (12)$$

Several statistics have been proposed for comparing these variance-covariance matrices. A suggested approach is based on analysis of the ratio of determinants of both matrices. If the effect of the deletion of the ith observation from the model is minor, the ratio of the computed values of both determinants would be close to one. On the other hand, if the value of the ith observation is atypical, its deletion from the model, would result in a value of this ratio far from one.

A limitation in using this ratio is the fact that the estimator of σ given by S is also affected by the deletion of the ith observation. However, Belsley, Kuh and Welsch (1980) show that by forming the determinantal ratio of both matrices, i.e., with all, and with the 'not ith' observation, a test statistic results

$$\mathrm{COVRATIO} = \frac{s(i)^{2p}}{s^{2p}} \left\{ \frac{|(X^1(i)X(i))^{-1}|}{|(X^1X)^{-1}|} \right\} \qquad (13)$$

Values of this ratio outside the interval 1 ± 3p/n identify items whose deletion cause atypical perturbations on the estimates of the covariance-ratio. In summary, values of this determinantal ratio greater than one, imply that the deletion of the ith item impairs estimation efficiency. Conversely, determinantal ratios less than one imply that the deletion of ith item enhances estimation efficiency.

Values of the covariance ratio are recorded in column 4 of tables 1 and 2. Examination of these estimates indicates that the deletion of item 1 causes an unusually large perturbation on this statistic. Its computed value of .70 lies outside the interval (.80 - 1.20). This result is consistent with previous findings which identify item 1 as eliciting a pattern of influence statisticallly different from the remaining items. A similar analysis of estimates of this ratio listed in table 2 (model 2), identifies four items whose deletions cause unusually large perturbations and lie outside the interval (.80 - 1.20). These items are: item 1, 13, 14, and 21. All but item 21 have been previously identified as items whose pattern of influence needs further examination.

Another important regression diagnostic is derived from Analysing the effect of the deletion of the ith observation on the predictive performance of a regression model. This effect can be conveniently summarized by the DFFIT coefficient. Following results of Belsley et. al., (1980), this statistic can be estimated by

$$DFFIT_i \equiv \hat{Y}_i-\hat{Y}_i(i) = x_i\left[\hat{\beta}-\hat{\beta}(i)\right] = h_i e_i/1-h_i \qquad (14)$$

For purposes of scaling, this quantity is divided by an estimate of $\sigma\sqrt{h_i}$. This adjustment yields the statistic

$$DFFITS_i = \frac{\sqrt{h_i}\, e_i}{s(i)(1-h_i)} \qquad (15)$$

where σ has been estimated by S(i). Estimates of this coefficient are recorded in column 5 of tables 1 and 2. Values of this statistic larger than 2 * $\sqrt{(p/n)}$ ex ert atypical effects on the predictive performance of the model. The DFFIT statistic is useful in the following context. Outliers often pull the estimated regression plane towards themselves. This often results in residual values smaller than their true value. The DFFIT statistic avoids this problem by re-estimating each residual with regression estimates that do not use that observation. The DFFIT statistic offers a very sensitive regression diagnostic for detecting potentially biased items, by identifying unusual patterns of influence on the predictive ability of the model.

Another important regression diagnostic applied to detect potentially biased items is based on analysis of the magnitude of the changes on the regression coefficients caused by the deletion on each item. In the simple bivariate model, for example, items whose deletion effect large perturbation on the intercept and slope estimates can be readily identified. Their large effects on the regression coefficients may indicate particular characteristics of an item that is lacking in others. These characteristics may, in turn, either increase or decrease the a priori probability of a correct response in one group of examinees but not in another. The identification of items whose deletion cause large perturbations on estimates of the regression coefficients is therefore of great value in helping to detect potentially biased items. Atypical perturbations in estimates of regression coefficients that may ensue as a result of their deletion can greatly facilitate the identification of atypical items. If we let b(i) be the vector of regression coefficients in a model that does not use the ith observation, the change or sensitivity of these coefficients can be estimated by

$$\mathrm{DFBETAS}_{ij} = \left[\,\hat{\beta}_j - \hat{\beta}_j(i)\right] / \left[s(i)\sqrt{(X^1X)^{-1}_{\ ij}}\,\right] \tag{16}$$

Belsley et. al., (1980) suggest several statistical criteria to set cutoff levels to identify atypical coefficient changes. A proposed cutoff is $2 / \sqrt{n}$. This cutoff measures the change in the estimates of the regression coefficients in units measured in standard deviations. In our analysis, items whose deletion cause a change of a least .365 standard deviations are deemed influential and warrant further examination for potential bias. Items whose DFBETAS exceed this cutoff are noted in columns 6 and 7 of tables 1 and 2 respectively.

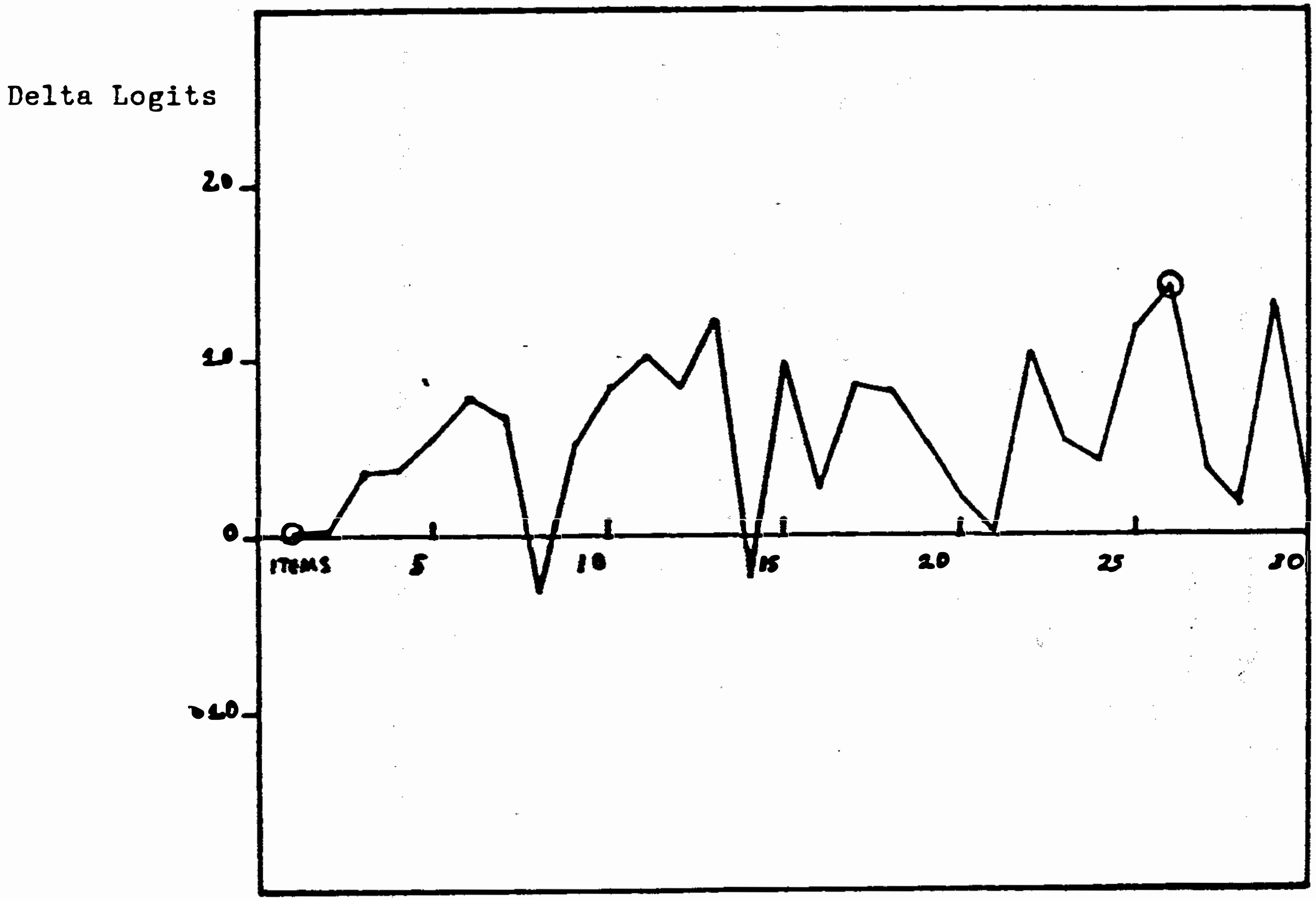
Further statistical analysis was carried out on the differences of logits of individual item p-values. These differences or delta values are defined as

$$\text{DELTA} = \text{LOGIT}(P_w) - \text{LOGIT}(P_b) \qquad (17)$$

A plot of these values against national P-values is given in figure 2. Under the assumption of equal performance, a fitted line through these values is expected to have a zero slope and zero intercept term. The observed dispersion of these DELTA values above zero indicates that a higher proportion of white examinees relative to black examinees has responded correctly to those exercises. The magnitude of these DELTA values is not,however,constant. From figure 2, a gradual increase in their magnitude is apparent. This trend suggests that the difference in performance between white and black examinees is not as marked among difficult items, as it is among relatively easier items. This performance

FIGURE 2

PLOT OF DELTA VALUES OF LOGITIZED P-VALUES



differential suggests that some items are equally difficult for both white and black examinees. However, as the level of difficulty decreases, a higher proportion of white examinees relative to black examinees succeeds in given a correct answer. A least squares fit to the dispersion of DELTA values produces a significant slope estimate (.01, p=.001). The estimate of the intercept term is not statistically different from zero (-.07,p=.63). From this gradual pattern in the magnitude of DELTA values, items that elicit atypical performance patterns can then be identified and contrasted with previous results.

Results of analysis of the regression diagnostics is listed in table 3. Examination of the magnitude of raw and studentized residuals identifies items 1 and 26 as eliciting residual values statistically different from the dispersion pattern established by the remaining items. This result is consistent with previous results, which identify the same items as atypical. Analysis of estimates of the covariance ratio identify items 1, 14 and 21 as exceeding the interval (.80 - 1.20). The extremely low value of this ratio due to the deletion of item 1 indicates that this item is highly atypical. This result contrasts well with our previous findings based on predictive models of white and black per-

TABLE 3

Delta Logits Regression Model

Model 3

Item No.	Hat Matrix	Raw Resid.	Stdzed. Resid.	Covar. Ratio	DFFITS	DFBETAS Const.	Slop
1	0.09	- .95	-3.44*	0.57*	-1.09*	0.52*	-0.8
2	0.03	- .49	-1.51	0.94	-0.28	-0.14	0.0
3	0.04	- .05	-0.15	1.12	-0.03	-0.02	0.0
4	0.07	- .53	-1.68	0.95	-0.49	0.20	-0.3
5	0.03	- .11	-0.32	1.10	-0.06	-0.00	-0.0
6	0.05	- .00	-0.02	1.13	-0.00	0.00	-0.0
7	0.04	0.27	0.82	1.07	0.17	0.14	-0.0
8	0.13	- .38	-1.21	1.11	-0.47	-0.47	0.4
9	0.04	0.14	0.42	1.11	0.09	0.08	-0.0
10	0.04	0.42	1.29	0.99	0.26	0.20	-0.1
11	0.06	0.16	0.48	1.12	0.12	-0.03	0.0
12	0.03	0.19	0.57	1.08	0.11	0.01	0.0
13	0.10	0.21	0.65	1.15	0.21	-0.11	0.1
14	0.15	- .25	-0.80	1.21*	-0.34	-0.33	0.
15	0.05	0.18	0.53	1.11	0.12	-0.03	0.
16	0.06	- .00	-0.00	1.14	-0.00	-0.00	0.
17	0.03	0.21	0.64	1.08	0.12	0.02	0.
18	0.04	.05	0.17	1.12	0.03	-0.00	0.
19	0.08	0.32	0.99	1.08	0.29	0.28	-0.
20	0.03	- .33	-0.99	1.03	-0.18	-0.07	-0.
21	0.12	- .00	-0.01	1.23*	-0.00	-0.00	0.
22	0.07	0.10	0.31	1.15	0.09	-0.04	0.
23	0.03	.01	0.03	1.11	0.00	0.00	-0.
24	0.07	0.15	0.46	1.14	0.13	0.12	-0.
25	0.04	0.43	1.32	0.99	0.28	-0.03	0.
26	0.06	0.55	1.73*	0.93	0.46	-0.16	0.
27	0.08	- .49	-1.53	0.98	-0.45	0.19	-0.
28	0.06	- .12	-0.37	1.13	-0.09	-0.09	0.
29	0.09	0.33	1.03	1.09	0.32	-0.15	0.
30	0.06	-0.05	-0.16	1.14	-0.04	-0.04	0.

formance. Similarly, analysis of the significance of the DFFITS and DFBETAS statistics consistently identifies item 1 as eliciting perturbations statistically different from those caused due to the deletion of the remaining items.

CONCLUSIONS

Results of applying the regression diagnostics proposed in this investigation consistently identify items 1 and 26 as eliciting response patterns statistically different from those observed in the remaining items. Although the preceding results do not imply that these items are biased, the magnitude of the perturbation on several statistics due to their deletion suggests that these items deem further examination.

Given the preceding, the performance of these two groups in these two items was further analyzed. Results of analysis of item 1 indicates that the performance of white and black examinees in this particular item was almost identical, with observed p-values of 93.6 and 93.5 respectively. This is a very atypical performance that substantially deviates from the pattern established by these groups of examinees in the remaining items.

By contradistinction, analysis of item 26 indicates that the observed performance gap is highly atypical. The observed p-values of 87.7 and 63.1 for white and black examinees respectively, substantially deviate from the distribution of performance values observed in the remaining items. Although the preceding results do not imply that these items are biased, the highly atypical performance levels they elicit among these examinees needs serious further examination. Item 26 in particular elicits an inordinately large performance gap that far exceeds the performance differential observed in the remaining items between black and white examinees.

The preceding results indicate how the recent developments in the analysis of regression models may prove useful in the identification of atypical and potentially biased items. Moreover, it is contended that the application of statistical criteria to set cutoff levels and identify atypical observations offers a substantial refinement over existing approaches, namely, delta plot , chi-square and latent trait methods.

FIGURE 3

STUDENTIZED RESIDUALS

WHITE REGRESSION MODEL

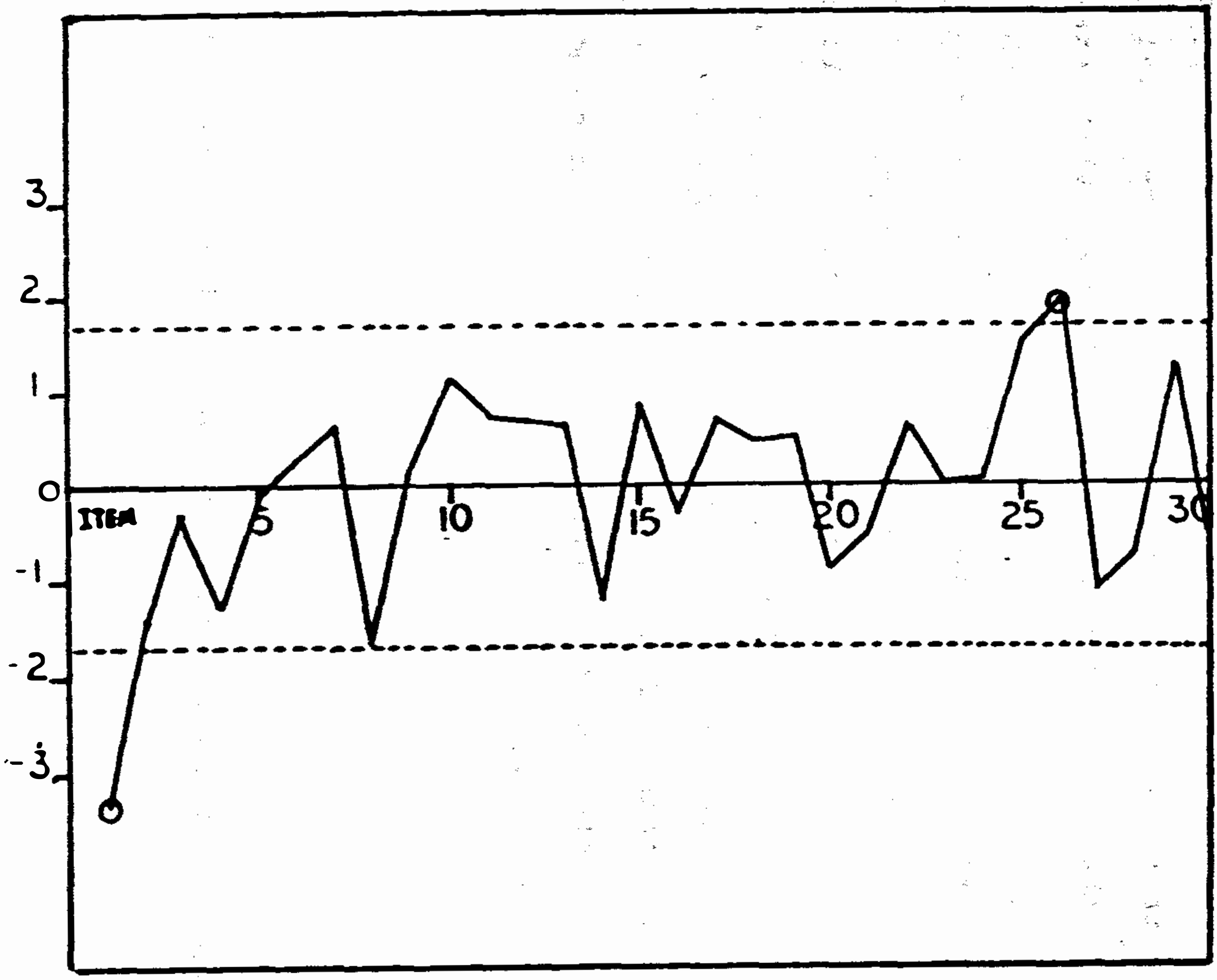


FIGURE 4

STUDENTIZED RESIDUALS

BLACK REGRESSION MODEL

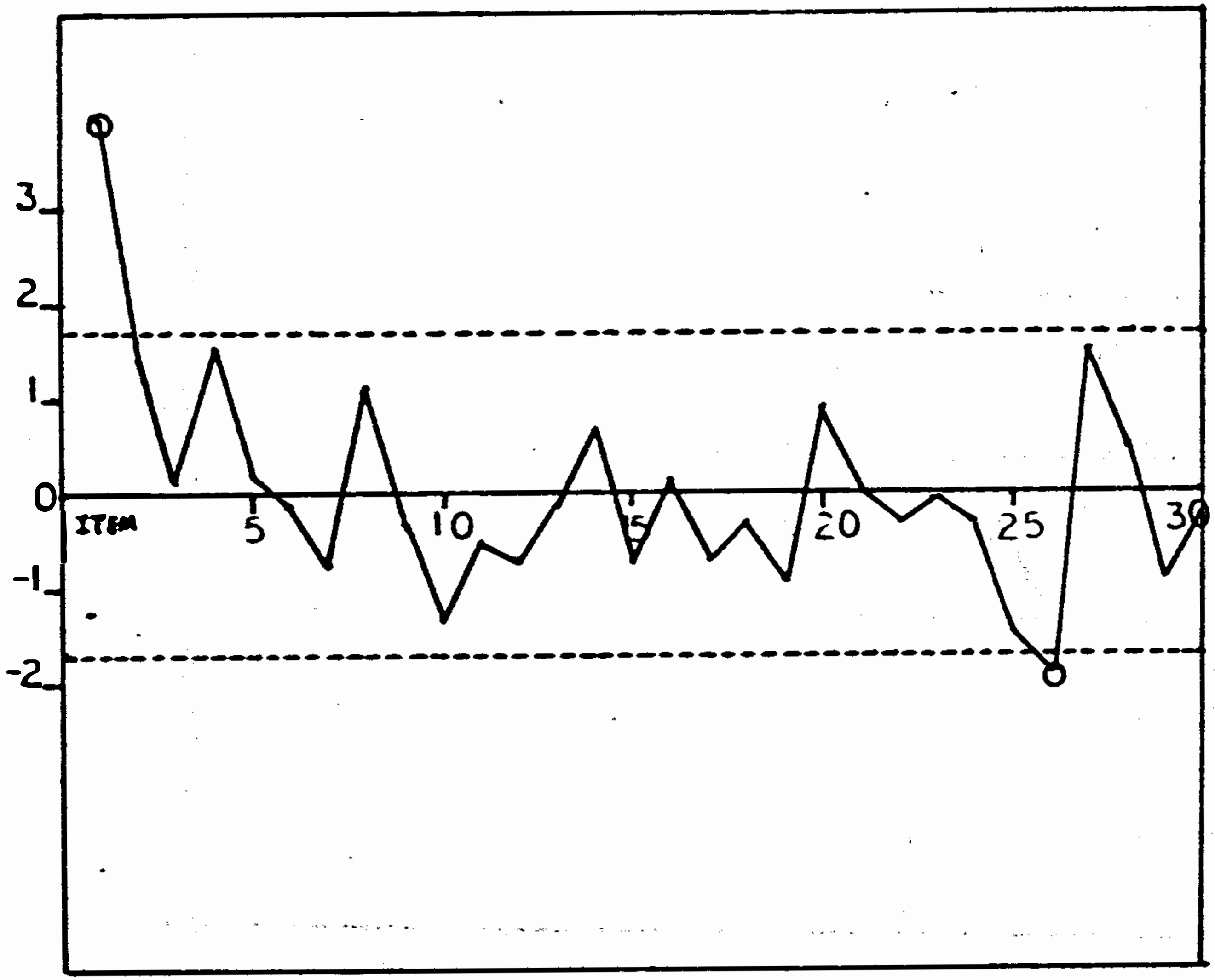
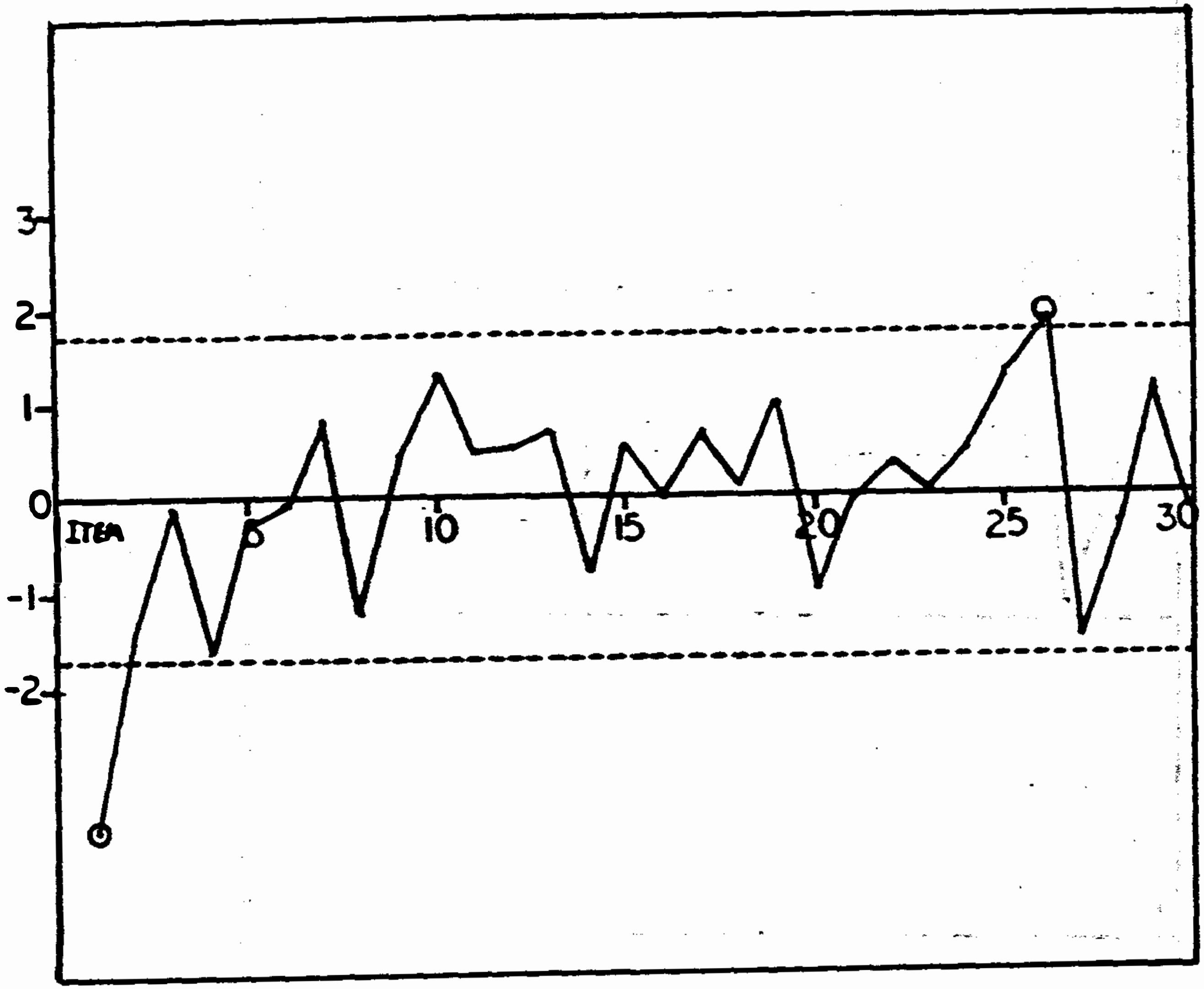


FIGURE 5

STUDENTIZED RESIDUALS

DELTA LOGITS REGRESSION MODEL



REFERENCES

Angoff, W. H., and Ford, S. F. Item-Race Interaction on a Test of Scholastic Aptitude, Journal of Educational Measurement, 1973, Vol. 10, pp. 95-105.

Belsley, D. A., et al. Regression Diagnostics, John Wiley and Sons: New York, 1980.

Camilli, G. A Critique of the Chi-Square Method for Assessing Item Bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder, 1979.

Hambleton, R. K. and Cook, L. L. Latent Trait Models and Their Use in the Analysis of Educational Test Data, Journal of Educational Measurement, 1977, Vol. 14, pp. 75-96.

Hoaglin D. C. and R. E. Welsch. The Hot Matrix on Regression and ANOVA, The American Statistician, 1978, 32, pp. 17-22.

Merz, W. R., et al. An Empirical Investigation of Six Methods for Examining Test Item Bias. Report submitted to the National Institute of Education, Grant 6-78-0067, California State University, Sacramento, California, 1978.

Lord, F. M. A Study of Item Bias Using Item Characteristic Curve Theory. In N. H. Poartinga (ed.) Basic Problems in Cross-cultural Psychology, Amsterdam: Swits and Vitlinger, 1977.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores, Addison-Wesley: Reading, Massachusetts, 1918.

Petersen, N. S. Bias in the Selection Rule: Bias in the Test. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977.

Rudner, L. M., Getson, P. R. and Knight, D.L. A Monte Carlo Comparison of Seven Biased Item Detection Techniques, Journal of Educational Measurement, 1980, Vol. 17, pp. 1-10.

Scheuneman, J. A Method for Assessing Bias in Test Items, Journal of Educational Measurement, 1979, Vol. 16, pp. 143-152.

Sheppard, L. A., et al. Comparison of Six Procedures for Detecting Test Item Bias Using Both Internal and External Ability Criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.