DOCUMENT RESUME

ED 281 883 TM 870 291

AUTHOR Seong, Tae-Je; Subkoviak, Michael J.
TITLE A Comparative Study of Recently Proposed Item Bias Detection Methods.
PUB DATE Apr 87
NOTE 33p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Blacks; Difficulty Level; Higher Education; *Item Analysis; Measurement Techniques; *Minority Groups; Multiple Choice Tests; Sample Size; *Test Bias; *Test Items; *Test Theory; Vocabulary Skills
IDENTIFIERS A Priori Tests; *Chi Square Analysis; Item Discrimination (Tests); Point Biserial Correlation; *Three Parameter Model

ABSTRACT

The purpose of this research was to reinvestigate the accuracy of three item bias detection procedures: (1) Linn and Harnisch's pseudo-IRT(Z) method; (2) Camilli's chi-square technique; and (3) Angoff's revised transformed item difficulty method. These methods are applied when the minority group sample size is too small to obtain stable estimates of item parameters. This study analyzed the data which included ten black slang items imbedded within a standardized vocabulary test. In order to determine the best methodology, three statistics were calculated: a point biserial correlation between an a priori bias index and the detected bias index associated with each method, intercorrelations among the bias measures of three procedures, and the percentage of agreement between the a priori bias index and bias index based on each method. Results showed that (1) the chi-square technique is slightly more accurate than the pseudo-IRT(Z) method in detecting bias; (2) Angoff's revised transformed item difficulty (TID) method is considerably worse; and (3) the chi-square procedure is highly correlated with the pseudo-IRT(Z) method. Appendices include item bias indices for all items and all methods, item information for computing the item bias index of Angoff's revised TID method for white and black groups, estimates of item parameters based on the three-parameter logistic model for Linn-Harnisch's method, and principal component analysis of the test item. (Author/JAZ)

A Comparative Study of Recently Proposed Item Bias Detection Methods

Seong, Tae-Je

and

Subkoviak, Michael J.

The University of Wisconsin-Madison

Paper presented at the Annual Meeting of the National Council on Measurement in Education

Washington, DC, April, 1987

ABSTRACT

The purpose of this research was to reinvestigate the accuracy of three item bias detection procedures (Linn and Harnisch's pseudo-IRT(Z) method, Camilli's chi-square technique, and Angoff's revised transformed item difficulty method) in the typical situation where the minority group has a small number of examinees. The current study analyzed the data which included ten black slang items imbedded within a standardized vocabulary test. In order to determine the best methodology among three procedures, this study calculated three statistics: a point biserial correlation between an á prior bias index and the detected bias index associated with each method, intercorrelations among the bias measures of three procedures, and the percentage of agreement between the á priori bias index and bias index based on each method.

This study found that 1) the chi-square technique is slightly more accurate than the pseudo-IRT(Z) method in detecting bias; 2) Angoff's revised TID method is considerably worse; and 3) the chi-square procedure is highly correlated with the pseudo-IRT(Z) method.

There are two reasons why the pseudo-IRT(Z) may be less accurate than the chi-square technique. One reason is that the estimates of item parameters for the pesudo-IRT(Z) procedure may be influenced by combining the minority group

with the majority group. Another reason may be violation of test unidimensionality assumed by the pseudo-IRT(Z) method.

BACKGROUND AND PURPOSE

In the last two decades, gender and race differences in test outcomes have been special topics of interest in the field of education. Jensen's assertion (1969) that there was a difference of one standard deviation in intelligence between blacks and whites continues to cause concern even today. Walker's study (1984) used meta analysis to check the widely shared assumption that women fixate at stage 3 in moral reasoning, and men progress to stage 4. Prior to discussing the argument that one group is better than another, it is important to investigate the question of whether test items are biased against certain subgroups. Williams (1971) insisted that traditional educational and employment tests are oriented toward the white middle class. Faggen-Steckler, McCarthy, and Title (1974) found that considerable content bias exists even in standardized tests in terms of the number of noun and pronoun references with respect to gender. As an indication that bias is an important topic, the Spring, 1976 issue of the Journal of Educational Measurement was devoted entirely to bias in selection.

Many psychometricians have attempted to provide a concrete and clear definition of item bias since the late 1960s. Cleary and Hilton (1968) defined bias as an interaction between item and group in terms of analysis of

variance. Angoff and Ford (1973) said that an item is considered biased if the item difficulty index or P-value for one group is relatively higher or lower than that for another group. Scheuneman (1975, 1979) stated that an item is biased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is different for various population subgroups being considered. A widely accepted definition is: An item is biased if individuals with equal ability, but from different groups, have unequal probability of answering the item correctly.

Shepard (1982) categorized approaches for detecting item bias, e.g., judgmental review, statistical review, and posterior analysis. Schmeiser also (1982) classified three approaches to detect item bias; these are the judgmental method, statistical item bias method, and experimental design method. Various statistical methodologies have been proposed for detecting item bias:

(1) analysis of variance (Cleary & Hilton, 1968);

(2) distractor response analysis (Veale & Foreman, 1975);

(3) transformed item difficulty methods (Angoff,1982; Angoff & Ford, 1973; Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Williams, 1985);

(4) chi-square methods (Camilli, 1979; Scheuneman, 1975, 1979);

(5) item response theory methods (Draba, 1978; Durovic, 1975; Levine, 1981; Levine, Wardrop, & Linn, 1982; Linn, Levine, Hasting, & Wardrop, 1981; Linn & Harnisch, 1981; Rudner,1977; Wright, Mead, & Draba, 1976);

(6) logit model methods (Mellenbergh, 1982; van der Flier, Mellenbergh, Ader, & Wijn, 1984).

These methodologies are different but are concerned with the same concept of bias. They produce somewhat different results because of theoretical and practical reasons. Therefore, many studies have been devoted to comparisons of these methods (Ironson, 1977; Ironson & Subkoviak, 1979; Merz & Grossen, 1979; Rudner & Convey, 1978; Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1985; Subkoviak, Mack, Ironson & Craig, 1984). The most widely accepted methods appear to be the transformed item difficulty approach (Angoff & Ford, 1973), the item characteristic curve procedures (Draba, 1978; Durovic,1975; Lord, 1977; Rudner, 1977), and the chi-square methods (Camilli, 1979; Scheuneman, 1975, 1979) which are similar in certain respects to the item characteristic curve method. Although comparative studies agree that the best procedure is the three-parameter item characteristic curve method, followed by the chi-square method, and then the transformed item difficulty method, most of these studies have not included recently revised or new methods.

Recently, a number of new or modified methods have been proposed for detecting item bias when the minority sample is small. One of the new methodologies is Linn and Harnisch's (1981), so called, pseudo-IRT(Z) technique. Another new methodology is Mellenberg's (1982), using a log linear model for three-way contingency table of test score categories, groups, and item responses. Another modified technique is Angoff's revised transformed item difficulty procedure (1982).

Shepard, Camilli, and Williams (1985) investigated Linn and Harnisch's pseudo-IRT(Z) and Angoff's revised transformed item difficulty method and compared them to other commonly used methods. Their study concluded that (1) the pseudo-IRT(Z) is the method of choice when the sample size of the minority group has 300 or fewer members; (2) the pseudo-IRT(Z) method is highly correlated with the widely accepted three-parameter item characteristic curve method; (3) the pseudo-IRT(Z) method is more accurate than a chi-square method at identifying bias; and (4) Angoff's revised transformed item difficulty procedure is considerably worse.

However, additional studies are needed to confirm the above claims. The present study is interested in evaluating the accuracy of Linn and Harnisch's method. A primary question which the present study will attempt to answer is: Did the pseudo-IRT(Z) method perform well in Shepard et

al.'s study (1985) because biased items were defined as such via large sample IRT analysis? It will also be possible to examine Camilli's chi-square method and Angoff's revised transformed item difficulty method.

METHOD

Three statistical item bias indices were computed in this research: Linn-Harnisch's pseudo IRT(Z) index based on the three-parameter logistic model, a chi-square statistic resulting from Camilli's method, and a distance from a point to major axis resulting from Angoff's revised TID method. Both signed and unsigned measures were computed.

Pseudo-IRT(Z). Linn and Harnish (1981) proposed an alternative to the three-parameter item response theory method when the minority group sample size is too small to obtain stable estimates of the item parameters.

This procedure estimates item discrimination, item difficulty, and guessing parameters based on the combined sample of minority and majority group examinees. $P_i(\theta_j)$, the probability that examinee j will answer item i correctly, is obtained by the following formula:

$$P_i(\theta_j) = c_i + (1 - c_i) \frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \qquad (1)$$

,where θ_j = examinee ability level;
α_i = item discrimination parameter;
β_i = item difficulty parameter; and
c_i = item guessing parameter.

Minority group examinees are divided into quintiles on the basis of their estimated ability levels. A standardized difference score for examinees in quintile q is then computed as follows:

$$Z_{iq} = \frac{1}{N_q} \sum_{j \in q} \frac{U_i(\theta_j) - P_i(\theta_j)}{\sqrt{P_i(\theta_j)Q_i(\theta_j)}} \qquad (2)$$

,where $U_i(\theta_j)$ = 1 if person j answers item i correctly or 0 otherwise;
$P_i(\theta_j)$ = the estimated probability that person j answers item i correctly based on the combined group;
$Q_i(\theta_j)$ = $1 - P_i(\theta_j)$; and
N_q = the number of examinees in a quintile q.

Z_{iq} is an index of the degree to which the observed performance for members of quintile q is better or worse than predicted by the model in Equation (1). The following formula likewise is used to obtain a standardized difference for the complete minority group as an index of bias:

$$Z_{i\cdot} = \frac{\sum_q N_q Z_{iq}}{\sum_q N_q} \qquad (3)$$

The $Z_{i\cdot}$ index will be 0 when an item is not biased; while a large $Z_{i\cdot}$ value (positive or negative) suggests the presence of bias. A positive sign indicates that an item favors the minority group, because their actual performance is better than their expected performance. Actually, a signed or unsigned $Z_{i\cdot}$ index can be calculated. If the direction of bias is not consistent across the quintiles, the signed

index is small. An unsigned Z_i. index is simply the sum of the absolute values of $N_q Z_{iq}$ in Equation (3).

Chi-square. Scheuneman (1975) said that "an item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered". Baker (1981) pointed out several problems with Scheuneman's procedure which focused only on correct responses to an item. Camilli (1979) modified the procedure to consider both correct and incorrect responses in an analysis.

The chi-square procedure divides the total test score scale into discrete ability intervals or score levels. The present study used five test score intervals because Rudner et al. (1980) found that the chi-square technique using five intervals was as effective as the three-parameter item characteristic curve method under most of the investigated conditions. Observed frequencies are counted within each interval for each group with regard to the correct and incorrect answer respectively. Expected frequencies are computed by multiplying the proportion of examinees who respond correctly or incorrectly to the item within a total score interval by the total number of examinees within the each score interval for each group.

The observed frequencies are compared with the expected frequencies using a type of chi-square statistics that is sum of terms $(O - E)^2 / E$, across all intervals and groups. The full chi-square statistics for all responses is the sum of the chi-square value for correct and incorrect response. The full chi-square statistics is the index of bias. A large value indicates greater bias. A signed measure can be computed by considering the direction of bias within each interval and by attaching a positive sign to the squared terms for the interval if the black group is favored and a negative sign if the white group is favored.

Revised TID. Angoff and Ford (1973) originally proposed a method based on the traditional item difficulty index. If an item is relatively more difficult in one subpopulation than another, it is considered as biased.

The item difficulty or P value (a proportion of subjects getting the item right) is computed separately for each group and for each item. The P value is transformed to a Z value which is the (1-P)th percentile of the standard normal distribution. After the transformation, the plot of Z values tends to be linear. A delta value is calculated from the Z value to eliminate negative values using the linear transformation: $\Delta = 4Z + 13$. A large Δ value indicates a difficult item. The pair of corresponding Δ values for each item is graphed on a two dimensional scatter plot for the

two groups. The plot of points appears in the form of an ellipse like the usual correlation diagram. The major axis line of the ellipse is drawn on the scatter diagram. A measure of bias is the length of the perpendicular line from a given point in the plot to the major axis. The formulas for determining the major axis and the perpendicular line are given in Angoff and Ford (1973). A large distance from the a given point for the item to the major axis indicates a more biased item.

Hunter (1975), Lord (1977), and Angoff (1982) have pointed out theoretical limitations of the transformed item difficulty method (TID). If there is a large difference in average abilities between two groups, the TID method may indicate bias where none exists. Thus Angoff (1982) recently proposed dividing Z by the item-total correlation, the classical item discrimination, to obtain a Z' index.

P values and Z values are computed exactly like those based on the original transformed item difficulty method. Next the item discrimination, the point biserial correlation between an item and total score in each group, is computed. The newly derived Z' value is then calculated by dividing the Z value by the item-test correlation. Z' is essentially equivalent to the X_{50} difficulty index of latent trait theory if the normal ogive is fitted to the item response data (Baker, 1965).

Once Z' values are known, the same procedure to plot delta values, to draw the major axis, and to measure the perpendicular distance from a point to the major axis is like that of Angoff and Ford's original transformed item difficulty method. A large distance indicates a more biased item. Both signed and unsigned indices were computed in the present study. A positive sign was attached to bias index if the black group was favored and a negative sign if the white group was favored.

Data Source.

Data for this investigation are from a study by Subkoviak, Mack, Ironson, and Craig (1984). The purpose of using this data set, in which bias has been deliberately manipulated by including black vocabulary items, is to investigate how Linn and Harnish's pseudo-IRT(Z), Angoff's revised transformed item difficulty method, and Camilli's technique perform when the biased items are known externally. Specifically, these data consist of responses to 50 multiple-choice vocabulary test items, including 10 black slang items which were intentionally written by a black author to be biased against whites, independent of any statistical index of item bias. The other 40 items were drawn from the verbal section of the College Qualification Test which is an aptitude test constructed for college students. Four-option multiple-choice items were used, and

subjects were asked to choose the option which is a synonym for a given word. Black slang items were inserted randomly in each block of five items on the test. Directions for the test informed students that some of the words are standardized English, while others are slang. Further details of the data are provided by Subkoviak et al. (1984).

There were 1,022 whites and 1,008 blacks. In this study, data for all 1,022 whites but only 300 blacks were analyzed; since the methods of interest here are especially recommended when the minority group is small. The 300 blacks were selected randomly from the entire sample of 1,008 blacks.

Analysis

For this study, the Pearson correlation coefficient was used to investigate the accuracy of each bias detection method. The point-biserial correlation between the á priori bias index and the detected bias measure for each method indicates how well each method detects the items intentionally written to be biased. The ten slang items were coded (1) and the forty standard vocabulary items were coded (0) as an index of the á priori bias intentionally included in the test. In addition to this correlation, percentage agreement statistics were computed to determine the proportion of items which are classified as biased for a particular method. The agreement statistic (%) is the

proportion of correct classification; that is, the number of biased items detected which are black slang items plus the number of unbiased items which are not black slang items, divided by the total number of items.

The degree of correlation among bias detection methods was also computed. The correlation between the bias indices for two methods indicates how closely one method is associated with another method.

RESULTS

Detection of á Prior Bias

Correlation The resulting item bias indices for all methods are reported in Appendix A. Pearson correlations between the á priori bias index (zero-one coding) and the bias index for each procedure are shown in Table 1.

Table 1. Correlations between á Priori Bias and Detected Bias

Method	Unsigned Measure	Signed Measure
Pseudo-IRT(Z)	.710	.762
Camilli's χ^2	.691	.798
Revised TID	.345	.522

Separate correlations with á priori bias were calculated for both signed and unsigned indices of bias. The unsigned pseudo-IRT(Z) and Camilli's χ^2 measures correlated .710 and .691, respectively, with the known bias; whereas Angoff's revised TID correlated .345. Similarly, the signed pseudo-IRT(Z) and the chi-square measures correlated .762 and .798 with á priori bias ; whereas the revised TID correlated .522. For the signed measures, which are more consistent with the á priori index, Camilli's chi-square procedure has the highest correlation, followed closely by Linn-Harnisch's pseudo-IRT(Z) method, with Angoff's revised TID procedure last.

Shepard et al.'s (1985) study produced similar results to that of this study. However in their study, the signed pseudo-IRT(Z) produced the highest correlation with external bias, followed by Camilli's signed chi-square, and then the delta plot procedure. In their study, the signed pseudo-IRT(Z) measure and the signed chi-square index were correlated .62 and .59, respectively, with an á priori index based on ICC-3 analysis of their data, which may have favored the IRT(Z) procedure.

This study confirmed that the correlation between á priori bias and the signed measures for all methods were higher than the corresponding correlations based on unsigned measures (Subkoviak et al., 1984). This is rational because the standard deviation of signed measures is larger than

that of unsigned measures as indicated in Table 2 and because signed measures are directional like the á priori index used to compute the Pearson correlation coefficient for each method.

Table 2. Descriptive Statistics of Bias Index for Each Method

	Unsigned Measure			Signed Measure		
	IRT(Z)	Camilli	χ^2 Rev.TID	IRT(Z)	Camilli	χ^2 Rev.TID
Mean	.232	60.9	25.6	.055	13.0	.0
Stdev.	.178	87.1	39.5	.266	106.0	47.2

Percentage Agreement Additional information used to investigate the accuracy of each bias detection method is a percentage agreement or concordance between á priori bias and the bias detection by each method. Items detected as biased by each method are compared to the known bias. Contingency Table 3 shows the proportion of items which are detected as biased from each method.

As Table 3 indicates for the unsigned measure, the pseudo-IRT(Z) procedure and Camilli's chi-square had 92% agreement with á priori bias, whereas the revised TID had 76%. It may be noted that the unsigned measure of the revised TID method detected only four items as biased among the ten slang items. Thus it falsely identified six unbiased items as biased.

Table 3. Contingency Table and Percentage Agreement of Each Method for Detecting the Imbedded Bias in the Test

Unsigned Measure

	A Priori			A Priori			A Priori		
	B	NB		B	NB		B	NB	
IRT(Z) B	8	2	10	χ^2 B	8	2	10	Rev TID B	4	6	10
IRT(Z) NB	2	38	40	χ^2 NB	2	38	40	Rev TID NB	6	34	40
	10	40			10	40			10	40	
Agreement Percentage	92 %			92 %			76 %		
Phi	.75			.75			.25		

Signed Measure

	A Priori			A Priori			A Priori		
	B	NB		B	NB		B	NB	
IRT(Z) B	9	0	9	χ^2 B	10	0	10	Rev TID B	8	1	9
IRT(Z) NB	1	40	41	χ^2 NB	0	40		Rev TID NB	2	39	41
	10	40			10	40			10	40	
Agreement Percentage	98 %			100 %			94 %		
Phi	.937			1.0			.807		

The signed measure of Camilli's chi-square method detected all black slang items as biased for 100 % agreement, while Linn and Harnisch's pseudo-IRT(Z) method and Angoff's revised transformed item difficulty method achieved 98% and 94% respectively. Signed measures appeared more accurate in detecting bias than unsigned measures. For

the present data, Camilli's chi-square appeared to be the best method, followed by Linn-Harnisch's Pseudo-IRT(Z), and finally Angoff's revised TID method.

Agreement Among Methods Pearson correlations are one measure of how much one bias technique is related to that of another procedure. Intercorrelation among the bias indices of three methods are reported in Table 4. The correlations among item bias detection procedures were separately computed for signed and unsigned bias measures.

Table 4. Intercorrelations Among Bias Measures

	Unsigned Measure		Signed Measure	
	IRT(Z)	Camilli's χ^2	IRT(Z)	Camilli's χ^2
Camilli's χ^2	.901		.893	
Revised TID	.399	.354	.451	.497

Camilli's chi-square procedure is correlated highly with the pseudo-IRT(Z) method for both signed and unsigned measures (r=.901, .893) because both procedures use the same type of definition of bias and quintile groupings. For both signed and unsigned measures, Angoff's revised TID procedure is associated weakly with the pseudo-IRT(Z) method and Camilli's chi-square procedure. These results confirm that the revised TID technique is not consistent with other bias methods (Shepard et al., 1985).

CONCLUSION AND DISCUSSION

Signed measures detected á priori bias more precisely than unsigned measures for each method. Furthermore Linn and Harnisch's pseudo-IRT(Z) method and Camilli's chi-square procedure were better at detecting á priori bias than Angoff's revised TID method. For the intercorrelation among bias methods, Camilli's chi-square technique was highly correlated with Linn and Harnisch's pseudo-IRT(Z) method ($r \geq .893$). However, Angoff's revised transformed item difficulty procedure is only weakly associated with the other two methods ($r \leq .497$).

This study supports Shepard et al.'s study showing that there is high agreement between the pseudo-IRT(Z) and the simpler chi-square method and that Angoff's revised transformed item difficulty is not in close agreement with the other two. In other words, Angoff's revised TID procedure does not generally appear to be a good method to detect item bias. In Angoff's revised transformed item difficulty method, low test-item correlations resulted in extreme values of Z' and misleading bias indices in the present study (see Appendix B).

This study shows somewhat different results from Shepard et al.'s study (1985). The current study suggests that Linn and Harnisch's pseudo-IRT(Z) method may be

slightly less accurate than Camilli's chi-square procedure. There are several reasons for this.

One reason is that it may not be appropriate to fit the three-parameter item response model to combined minority and majority data. The black slang items have low item discriminations because there are many whites and a small number of blacks in the data set. The estimates of item parameters may be influenced by the target group combined with the majority group (see Appendix C).

Another reason may be violation of test unidimensionality assumed by the Pseudo-IRT(Z) method. There are thirteen principal components in the test having eigen values greater than one (see Appendix D); but the scree plot of these eigen values suggests two (or possibly three) factors in the test. Only two of the unrotated factors have many items whose factor loadings exceed .35 in absolute value. The first unrotated factor is related to thirty-three standard vocabulary items, the second factor is related to five black slang items. Only one or two items account for the remaining factors. After varimax rotation of the two factor solution, the primary factor might be called standard vocabulary and the second factor black slang based on loadings exceeding .35.

Even though Linn and Harnisch's pseudo-IRT(Z) appears to be on theoretically sound ground because it retains the benefits of item response theory, it is questionable that

Linn and Harnisch's pseudo-IRT(Z) method is necessary better than Camilli's chi-square method when the minority group is small.

Referecne

ANGOFF,W.H. (1982). Use of item difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD:John Hopkins University Press, 96-116.

ANGOFF,W.H., & FORD,S.F.(1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

BAKER,F.B. (1965). Origins of the item parameters X_{50} and β as a modern analyses technique. Journal of Educational Measurement, 2, 167-180.

BAKER,F.B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

CAMILLI, G. (1979). A critique of the chi-square method assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

CLEARY,T.A., & HILTON,T.L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

DRABA,R.E. (1978). The Rasch model and legal criterion of a "reasonable" classification. Unpublished doctoral dissertation, University of Chicago.

DUROVIC,J.J. (1975). Definition of test bias: A taxonomy and illustration of an alternative model. Unpublished doctoral dissertation, State University of New York at Albany.

FAGGEN-STECKLER,J., McCARTHY,K.A., & TITTLE,C.K. (1974). A quantitative method for measuring sex bias in standardized tests. Journal of Educational Measurement, 11, 151-161.

HUNTER,J.E. (1975). A critical analysis of the use of item meansand item-test correlation to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

IRONSON,G.H. (1977). A comparative study of several methods of assessing item bias. Unpublished dissertation, University of Wisconsin-Madison.

IRONSON,G.H. (1982). Use of Chi-square and Latent Trait Approaches for Detecting Item Bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD:John Hopkins University Press, 96-116.

INRONSON,G.H. & SUBKOVIAK,M (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

JENSEN,A,R. (1969). How much can we boost IQ and scholastic achievement?. Cambridge: Harvard Educational Review, 39, 81-83.

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.

LEVINE,M.V. (1981). Weighted item bias statistics. Report 81-5, Urbana-Champaign, IL: Department of Educational Psychology, University of Illinois.

LEVINE,M.V., WARDROP,J.L., & LINN,R.L. (1982). Weighted mean square item bias statistics. Paper presented at the Annual Meeting of American Educational Research Association, New York.

LINN,R.L., LEVINE,M.V., HASTING,C.N. & WARDROP,J.L. (1981) Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

LINN,R.L. & HARNISCH,D.L. (1981). Interaction between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

LORD,F.M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

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

MERZ,W.R. & GROSSEN,N.E. (1979). An empirical investigation of six methods for examining test item bias. Report submitted to the National Institute of Education, Grant NIE-6-78-0067, California State University, Sacramento, CA.

OSTERLIND,S.J. (1983). Test item bias. Beveraly Hills: Sage publications.

RUDNER,L.M. (1977). An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University of America, Washington, D.C.

RUDNER,L.M., & CONVEY,J.L. (1978). An evaluation of select approaches for biased item identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto. (ERIC Document Reproduction Service No. ED 157942)

RUDNER,L.M., GETSON,P.R. & KNIGHT,D.L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10

SCHEUNEMAN,J.D. (1975). A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association. Washington, D.C.

SCHEUNEMAM,J.D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

SCHMEISER,C.B. (1982). Use of experimental design in statistical item bias studies. In R.A.Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD:John Hopkins University Press, 64-95.

SHEPARD,L.A. (1982). Definitions of item bias. In R.A.Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD:John Hopkins University Press, 9-30.

SHEPARD,L.A., CAMILLI,G., & AVERILL,M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

SHEPARD,L., CAMILLI,G. & WILLIAMS,D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105

SUBKOVIAK,M.J., MACK,J.S., IRONSON,G.H., & CRAIG,R.D. (1984). Empirical comparison of selected item detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

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.

VEALE,J.R. & FORMAN,D.I. (1975). Cultural validity of items and tests: A new approach. Score Technical Report. Iowa city, IA: Westinghouse Learning Corporation/ Measurement Research Center.

WALKER, L. J. (1984). Sex difference on the development of moral reasoning: A critical review. Child Development, 55, 677-691.

WILLIAMS,R.L. (1971). Abuse and misuse of testing black children. The Counseling Psychologist, 2, 62-73.

WOOD,R.L., WINGERSKY,M.S., & LORD,F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (ETS RM 76-6). Princeton, NJ: Educational Testing Service, 1976.

WRIGHT,B.D., MEAD,R.J., & DRABA,R. (1976). Detecting and correcting test item bias with a logistic response model. Research Memorandum NO. 22. Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Appendix A Item Bias Indices for All Items and All Methods

Item	US(Z)	S(Z)	US(χ^2)	S(χ^2)	US(D)	S(D)
1	0.097	0.001	48.213	-48.042	10.900	-10.900
2	0.131	-0.100	61.789	-61.714	12.226	-12.226
* 3	0.490	0.490	119.140	119.140	3.547	3.547
4	0.250	-0.113	24.475	-24.297	11.431	-11.431
5	0.128	0.128	29.281	-29.281	16.242	-16.242
6	0.195	0.168	33.869	-33.166	7.091	-7.091
* 7	0.119	-0.112	19.417	17.484	0.017	-0.017
8	0.240	0.171	18.091	-18.052	21.935	-21.935
9	0.099	0.099	18.386	13.677	38.119	-38.119
10	0.153	0.009	19.964	-8.440	10.361	-10.361
11	0.115	-0.020	18.744	-18.731	11.516	-11.516
12	0.139	-0.009	15.521	-15.521	10.929	-10.929
13	0.081	-0.023	44.753	-44.694	11.439	-11.439
14	0.092	-0.042	45.069	-44.994	11.328	-11.328
* 15	0.230	0.230	36.161	36.161	5.071	-5.071
16	0.144	0.068	21.424	-21.396	10.882	10.882
17	0.085	-0.080	2.579	-1.231	32.243	-32.243
18	0.090	-0.039	24.899	-23.623	22.496	-22.496
19	0.233	-0.164	29.399	-27.417	14.659	-14.659
* 20	0.751	0.751	258.060	258.060	100.220	100.220
21	0.251	-0.108	46.225	-46.046	14.040	-14.040
22	0.101	-0.085	20.400	-18.601	20.964	-20.964
23	0.285	0.125	26.351	-26.351	15.630	-15.630
* 24	0.444	0.444	207.350	207.350	14.708	14.708
25	0.163	-0.087	73.399	-73.361	16.432	-16.432
26	0.140	-0.026	4.225	2.301	13.565	-13.565
* 27	0.350	0.322	123.900	123.630	3.056	3.056
28	0.170	-0.170	46.837	46.684	13.186	-13.186
29	0.354	-0.354	47.198	-47.198	25.968	-25.968
30	0.107	0.008	7.966	-3.358	11.998	-11.998
31	0.249	-0.111	23.841	-23.841	12.339	-12.339
32	0.115	-0.115	45.085	-41.136	12.493	-12.493
33	0.127	0.000	28.652	-28.652	11.342	-11.342
* 34	0.299	0.299	76.942	76.942	230.107	230.107
35	0.194	-0.016	21.044	-20.890	11.982	-11.982
36	0.220	-0.220	37.822	-37.822	7.860	-7.860
* 37	0.829	0.829	375.290	375.290	120.699	120.699
38	0.142	0.086	16.689	-14.632	9.256	-9.256
39	0.104	0.008	1.856	1.827	10.078	-10.078
40	0.141	0.003	23.551	-10.900	72.561	-72.561
41	0.268	-0.268	14.731	-13.934	15.176	-15.176
42	0.193	0.193	18.339	-17.869	20.176	-20.176
* 43	0.463	0.463	154.900	154.764	3.717	3.717
44	0.144	-0.043	5.774	0.134	10.864	-10.864
45	0.268	-0.183	22.869	-22.227	11.750	-11.750
46	0.087	-0.023	31.151	-30.173	10.781	-10.781
47	0.306	-0.306	54.816	-54.816	118.842	118.842
48	0.180	-0.180	29.447	-29.372	16.261	-16.261
49	0.202	-0.202	141.110	-141.084	15.484	-15.484
* 50	0.851	0.851	429.400	429.400	44.176	44.176

* Black Slang Item

Appendix B Item Information for Computing the Item Bias Index of Angoff's Revised Transformed Item Difficulty Method

White Group

Item	freq.	correct	P	1-P	Z	r	Z'	Δ'
1	1022	874	0.855186	0.144814	-1.0589	0.348	-3.0428	0.829
2	1020	806	0.790196	0.209804	-0.8071	0.467	-1.7283	6.087
* 3	1000	519	0.519000	0.481000	-0.0476	0.087	-0.5471	10.811
4	1013	700	0.691017	0.308983	-0.4986	0.433	-1.1515	8.394
5	1022	991	0.969667	0.030333	-1.8759	0.190	-9.8732	-26.493
6	1022	842	0.823875	0.176125	-0.9302	0.413	-2.2523	3.991
* 7	1016	86	0.084646	0.915354	1.3745	0.171	8.0380	45.152
8	1022	978	0.956947	0.043053	-1.7159	0.195	-8.7995	-22.198
9	1014	266	0.262327	0.737673	0.6362	0.485	1.3118	18.247
10	1022	795	0.777887	0.222113	-0.7651	0.437	-1.7508	5.997
11	1012	731	0.722332	0.277668	-0.7465	0.378	-1.9749	5.101
12	1016	770	0.757874	0.242126	-0.6995	0.439	-1.5934	6.626
13	1018	802	0.787819	0.212181	-0.7989	0.461	-1.7330	6.068
14	1014	826	0.814596	0.185404	-0.8950	0.309	-2.8964	1.414
* 15	1016	594	0.584646	0.415354	-0.2138	0.392	-0.5454	10.818
16	1022	800	0.782779	0.217221	0.7816	0.480	-1.6283	6.487
17	1008	190	0.188492	0.811508	0.8835	0.282	3.1330	25.532
18	1019	387	0.379784	0.620216	0.3060	0.547	0.5594	15.238
19	1021	435	0.426053	0.573947	0.1864	0.596	0.3128	14.251
* 20	1015	443	0.436453	0.563547	0.1600	0.018	8.8889	48.556
21	1019	841	0.825319	0.174681	0.9358	0.315	-2.9708	1.117
22	1017	376	0.369715	0.630285	0.3326	0.594	0.5599	15.240
23	1021	940	0.920666	0.079334	-1.4096	0.313	-4.5035	-5.014
* 24	1004	366	0.364542	0.635458	0.3463	0.114	3.0377	25.151
25	1021	925	0.905975	0.094025	-1.3164	0.320	-4.1138	-3.455
26	1019	438	0.429833	0.570167	0.1768	0.471	0.3754	14.501
* 27	1005	390	0.388060	0.611940	0.2844	0.118	2.4102	22.641
28	1008	678	0.672619	0.327381	-0.4471	0.497	-0.8996	9.402
29	1018	555	0.545187	0.454813	-0.1133	0.285	-0.3975	11.410
30	1015	524	0.516256	0.483744	-0.0407	0.560	-0.0727	12.709
31	1022	915	0.895303	0.104697	-1.2552	0.425	-2.9534	1.186
32	1021	811	0.794319	0.205681	-0.8215	0.379	-2.1675	4.330
33	1020	838	0.821569	0.178431	-0.9213	0.451	-2.0428	4.829
* 34	993	25	0.025176	0.974824	1.9570	0.021	93.1905	385.762
35	1021	821	0.804114	0.195886	-0.8564	0.333	-2.5718	2.713
36	1021	657	0.643487	0.356513	0.3130	0.323	0.9690	16.876
* 37	1005	226	0.224876	0.775124	0.7558	0.021	35.9905	156.962
38	1020	839	0.822549	0.177451	-0.9251	0.409	-2.2619	3.953
39	1009	339	0.335976	0.664024	0.4234	0.226	1.8735	20.494
40	1015	320	0.315271	0.684729	0.4809	0.541	0.8889	16.556
41	1015	403	0.397044	0.602956	0.2610	0.351	0.7436	15.974
42	1021	989	0.968658	0.031342	-1.8614	0.193	-9.6446	25.578
* 43	1004	431	0.429283	0.570717	0.1782	0.218	0.8174	16.270
44	1018	538	0.528487	0.471513	-0.0714	0.473	-0.1510	12.396
45	1020	651	0.638235	0.361765	-0.3537	0.425	-0.8322	9.671
46	1017	727	0.714848	0.285152	-0.5674	0.439	-1.2925	7.830
47	1016	374	0.368110	0.631890	0.3369	0.425	0.7927	16.171
48	1021	559	0.547502	0.452498	-0.1194	0.483	-0.2472	12.011
49	1021	915	0.896180	0.103820	-1.2601	0.405	3.1114	0.555
* 50	998	195	0.195391	0.804609	0.8582	0.062	13.8419	68.368

* Black Slang Item

Black Group

Item	freq.	correct	P	1-P	Z	r	Z'	Δ'
1	295	180	0.610169	0.389831	-0.2798	0.206	-1.3583	7.567
2	299	135	0.451505	0.548495	0.1219	0.384	0.3174	14.270
* 3	299	251	0.839465	0.160535	-0.9924	0.250	-3.9696	-2.878
4	296	119	0.402027	0.597973	0.2481	0.426	0.5824	15.330
5	300	255	0.850000	0.150000	-1.0364	0.176	-5.8886	-10.555
6	295	205	0.694915	0.305085	-0.5097	0.265	-1.9234	5.306
* 7	298	32	0.107383	0.892617	1.2405	0.237	5.2342	33.937
8	300	251	0.836667	0.163333	-0.9810	0.333	-2.9459	1.216
9	294	61	0.207483	0.792517	0.8151	0.068	11.9868	60.947
10	299	164	0.548495	0.451505	-0.1219	0.358	-0.3405	11.638
11	296	154	0.520270	0.479730	-0.0509	0.330	-0.1542	12.383
12	296	148	0.500000	0.500000	0.0000	0.333	0.0000	13.000
13	299	148	0.494983	0.505017	0.0125	0.282	0.0443	13.177
* 14	295	172	0.583051	0.416949	-0.2098	0.195	-1.0759	8.696
* 15	298	188	0.630872	0.369128	-0.3342	0.326	-1.0252	8.899
16	299	152	0.508361	0.491639	-0.0211	0.434	-0.0486	12.806
17	291	34	0.116838	0.883162	1.1908	0.102	11.6745	59.698
18	292	52	0.178082	0.821918	0.9226	0.155	5.9523	36.809
19	300	35	0.116667	0.883333	1.1913	0.391	3.0468	25.187
* 20	300	290	0.966667	0.033333	-1.8330	0.065	-28.2000	-99.800
21	298	159	0.533557	0.466443	-0.0842	0.384	-0.2193	12.123
22	296	35	0.118243	0.881757	1.1837	0.218	5.4298	34.719
23	297	211	0.710438	0.289562	-0.5544	0.503	-1.1022	8.591
* 24	298	243	0.815436	0.184564	-0.8980	0.202	-4.4455	-4.782
25	299	174	0.581940	0.418060	-0.2068	0.444	-0.4658	11.137
26	297	80	0.269360	0.730640	0.6146	0.225	2.7316	23.926
* 27	300	200	0.666667	0.333333	-0.4308	0.410	-1.0507	8.797
28	292	96	0.328767	0.671233	0.4432	0.313	1.4160	18.664
29	297	76	0.255892	0.744108	0.6560	0.105	6.2476	37.990
30	296	90	0.304054	0.695946	0.5125	0.288	1.7795	20.118
31	300	196	0.653333	0.346667	-0.3942	0.503	-0.7837	9.865
32	300	150	0.500000	0.500000	0.0000	0.253	0.0000	13.000
33	299	163	0.545151	0.454849	-0.1133	0.409	-0.2770	11.892
* 34	292	34	0.116438	0.883562	1.1928	0.203	5.8759	36.503
35	300	175	0.583333	0.416667	-0.2103	0.382	-0.5505	10.798
36	294	109	0.370748	0.629252	0.3300	0.247	1.3360	18.344
* 37	296	257	0.868243	0.131757	-1.1178	0.112	-9.9804	-26.921
38	297	186	0.626263	0.373737	-0.3221	0.270	-1.1930	8.228
39	295	86	0.291525	0.708475	0.5488	0.187	2.9348	24.739
40	298	54	0.181208	0.818792	0.9108	0.039	23.3538	106.415
41	292	53	0.181507	0.818493	0.9097	0.251	3.6243	27.497
42	296	257	0.868243	0.131757	1.1178	0.258	-4.3326	-4.330
* 43	299	231	0.772575	0.227425	-0.7474	0.271	-2.7579	1.968
44	293	106	0.361775	0.638225	0.3536	0.268	1.3194	18.278
45	296	104	0.351351	0.648649	0.3815	0.386	0.9883	16.953
46	295	141	0.477966	0.522034	0.0551	0.240	0.2296	13.918
47	292	26	0.089041	0.910959	1.3469	-0.032	-42.0906	-155.362
48	296	73	0.246622	0.753378	0.6852	0.223	3.0726	25.291
49	300	142	0.473333	0.526667	0.0669	0.467	0.1433	13.573
* 50	295	253	0.857627	0.142373	1.0695	0.240	-4.4563	-4.825

* Black Slang Item

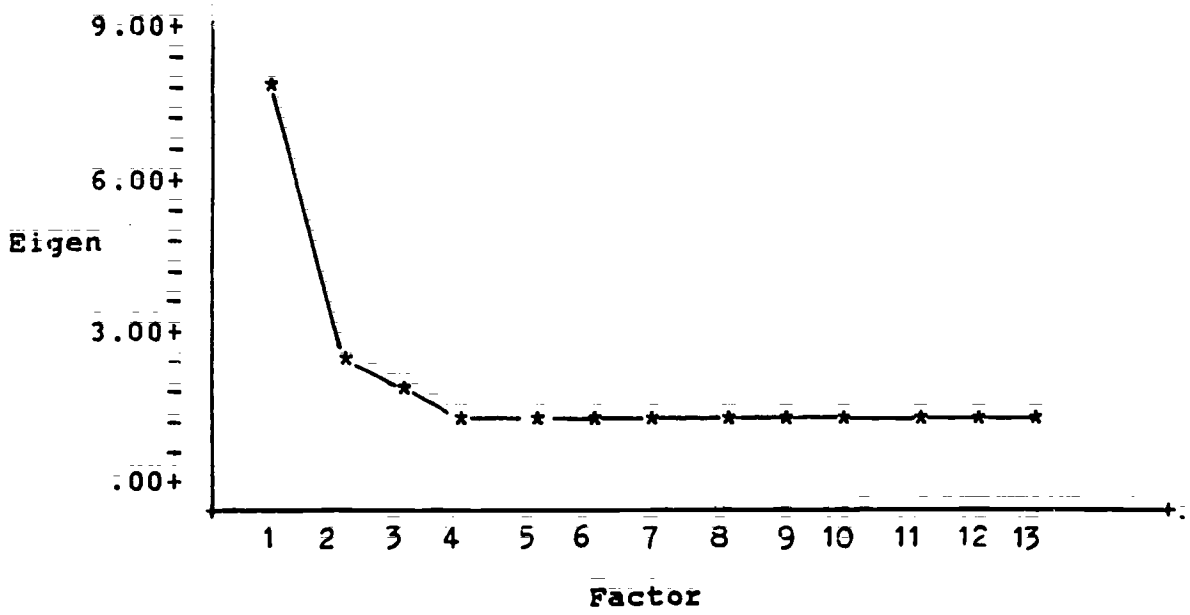
Appendix C Estimates of Item Parameters Based on the Three-Parameter Logistic Model for Linn-Harnisch's Pseudo-IRT(Z) Method

Item	a_i	b_i	c_i
1	0.8560	-1.0109	0.2000
2	1.3384	-0.4771	0.2000
*3	0.0100	-0.6538	0.2000
4	1.0361	-0.1172	0.2000
5	0.7719	-2.4935	0.2000
6	0.7955	-1.0381	0.2000
*7	0.3473	5.3703	0.0468
8	0.8304	-2.1954	0.2000
9	2.0183	1.3524	0.1500
10	0.9364	-0.6119	0.2000
11	0.6793	-0.4031	0.2000
12	1.0689	-0.4166	0.2000
13	1.2500	-0.4930	0.2000
14	0.6159	-0.9749	0.2000
*15	0.4271	0.0635	0.2000
16	1.7185	-0.4623	0.2000
17	1.1328	2.0892	0.1081
18	2.0183	0.9116	0.1503
19	1.4022	0.5090	0.0270
*20	0.0100	-0.3556	0.2000
21	0.7466	-0.8769	0.2000
22	1.9398	0.7739	0.0833
23	1.4341	-1.2242	0.2000
*24	0.0100	0.1596	0.2000
25	1.2097	-1.0692	0.2000
26	1.4182	0.9048	0.2000
*27	0.0216	21.3030	0.2000
28	1.2732	0.0131	0.2000
29	0.5304	0.8238	0.2000
30	1.7878	0.4708	0.2000
31	1.9511	0.9734	0.2000
32	0.8314	-0.6509	0.2000
33	1.2749	-0.6806	0.2000
*34	2.0183	4.2778	0.0507
35	0.6953	-0.9002	0.2000
36	0.6149	0.1182	0.2000
*37	0.0195	37.9377	0.2000
38	0.9211	-0.8608	0.2000
39	2.0183	1.9459	0.2816
40	2.0183	1.0897	0.1500
41	0.7440	1.2342	0.1500
42	0.9055	-2.2820	0.2000
*43	0.0140	19.8862	0.2000
44	0.9351	0.5153	0.2000
45	0.8875	0.1071	0.2000
46	0.9206	-0.2600	0.2000
47	1.3728	1.1466	0.1500
48	1.3651	0.4872	0.2000
49	1.7146	-0.8213	0.2000
*50	0.0100	86.3436	0.2000

* Black Slang Item

Appendix D Principal Component Analysis of the Test Item

Unrotated Factor	Eigen Value	NO.of Standard Items	NO.of Slang Items	NO.of Total Items
1	8.044	33	1	34
2	2.320	2	5	7
3	1.961	2	1	3
4	1.217	1	0	1
5	1.192	0	1	1
6	1.169	0	0	0
7	1.124	0	1	1
8	1.109	1	1	2
9	1.088	1	0	1
10	1.075	0	0	0
11	1.041	2	0	2
12	1.011	1	1	2
13	1.001	0	0	0



Unrotated Loadings for First Three Factors

		Unrotated Factor Loading		
		1	2	3
	ITEM(19)	0.614	0.207	-0.161
	ITEM(22)	0.575	0.274	-0.233
	ITEM(49)	0.546	-0.233	0.184
	ITEM(28)	0.536	-0.004	-0.041
	ITEM(30)	0.524	0.164	-0.091
	ITEM(2)	0.515	-0.044	0.141
	ITEM(18)	0.513	0.284	-0.222
	ITEM(31)	0.513	-0.159	0.352
	ITEM(16)	0.508	-0.053	0.170
	ITEM(39)	0.156	0.097	-0.042
*	ITEM(34)	-0.098	0.253	0.225
	ITEM(41)	0.364	0.089	-0.142
	ITEM(17)	0.255	0.177	-0.187
	ITEM(5)	0.273	-0.218	0.111
	ITEM(45)	0.468	0.072	-0.020
	ITEM(29)	0.311	-0.060	-0.183
	ITEM(46)	0.440	0.038	0.033
	ITEM(6)	0.392	0.057	0.133
	ITEM(25)	0.431	-0.143	0.144
*	ITEM(43)	-0.068	0.409	0.236
	ITEM(26)	0.418	0.266	-0.149
	ITEM(33)	0.499	-0.030	0.175
	ITEM(38)	0.407	-0.033	0.165
*	ITEM(7)	0.107	0.254	0.121
	ITEM(12)	0.465	0.030	0.126
*	ITEM(20)	-0.316	0.407	0.161
	ITEM(36)	0.359	0.021	-0.040
	ITEM(9)	0.365	0.392	-0.250
	ITEM(21)	0.402	-0.113	0.127
*	ITEM(37)	-0.348	0.462	0.272
	ITEM(48)	0.480	0.120	-0.108
*	ITEM(27)	-0.063	0.281	0.318
	ITEM(23)	0.442	-0.191	0.346
	ITEM(44)	0.418	0.190	0.014
	ITEM(8)	0.285	-0.150	0.364
	ITEM(32)	0.426	-0.073	0.092
*	ITEM(24)	-0.184	0.376	0.306
	ITEM(47)	0.442	0.109	-0.324
	ITEM(35)	0.382	-0.097	0.198
*	ITEM(50)	-0.351	0.494	0.263
	ITEM(10)	0.468	-0.027	0.145
*	ITEM(15)	0.248	0.323	0.266
	ITEM(1)	0.410	-0.076	-0.073
	ITEM(40)	0.426	0.387	-0.310
	ITEM(4)	0.491	0.006	0.023
	ITEM(11)	0.395	0.068	0.018
	ITEM(13)	0.486	-0.046	0.075
*	ITEM(3)	-0.122	0.223	0.389
	ITEM(42)	0.235	-0.189	0.193
	ITEM(14)	0.328	-0.107	0.077