DOCUMENT RESUME

ED 206 650 TM 810 540

AUTHOR Osterlind, Steven J.; Martois, John S.
TITLE Latent Trait Theory Applications to Test Item Bias Methodology. Research Memorandum No. 1.
INSTITUTION Los Angeles County Superintendent of Schools, Calif. Div. of Program Evaluation, Research, and Pupil Services.; Oakland Unified School District, Calif. Dept. of Research and Evaluation.
PUB DATE Dec 80
NOTE 35p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; High Schools; *Latent Trait Theory; Racial Bias; *Statistical Analysis; *Test Bias; Test Items
IDENTIFIERS *Rasch Model

ABSTRACT

This study discusses latent trait theory applications to test item bias methodology. A real data set is used in describing the rationale and application of the Rasch probabilistic model item calibrations across various ethnic group populations. A high school graduation proficiency test covering reading comprehension, writing mechanics, and mathematics was administered to 1,042 white and 11,441 black students in a large west coast school district. Using UCON estimation procedures for item difficulties, item plots for each ethnic group by the three separate subtests were prepared. The derivation of acceptable tolerance limits is described and applied to the current data set, wherein a biased item is revealed. The mathematics are given although their derivation is not described except when required for completeness. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206650

LATENT TRAIT THEORY APPLICATIONS TO TEST ITEM BIAS METHODOLOGY

by

Steven J. Osterlind
and
John S. Martois

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ The document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

Research Memorandum No. 1

Department of Research and Evaluation
Oakland Unified School District
and
Division of Program Evaluation, Research, and Pupil Services
Los Angeles County Superintendent of Schools

December 1980

TM 810 540

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Osterlind, S.
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

This study discusses latent trait theory applications to test item bias methodology. A real data set is used in describing the rational and application of the Rasch probabalistic model item calibrations across various ethnic group populations. The mathematics are given although their derivation is not described except when required for completeness. Using UCON estimation procedures for item difficulties item plots for each ethnic group by the several tests available (Reading, Written Expression, Mathematics) were prepared. The derivation of acceptable tolerance limits is described and applied to the current data set wherein a bias item is revealed.

LATENT TRAIT THEORY APPLICATIONS TO TEST ITEM BIAS METHODOLOGY

Unbiased student assessment on standardized tests currently in use is a quest fraught with confusion, misunderstanding, and misinterpretation in the current glass darkly debate over bias in mental testing. Issues raised are: 1) predictive validity for children from minority group backgrounds may be misrepresented by the standardization and validation groups; 2) for internal and construct criteria of bias, statistical adjustments alone (viz., regression techniques or ANOVAs regardless of how sophisticated) neither is supported by the empirical data available nor will likely gain acceptance (outside of psychometric debate) for administrative, political, and legal arguments; and, 3) as the cultural ethic changes from demands for equal opportunity to expectations of undifferentiated outcomes the discussion of differential validity and test bias will likely become more heated. Any resolution of these major points, of course, will hardly exhaust the argument (cf. Lord, 1971; Jensen, 1980; Scheuneman, 1975); and, as Pincher (1975) points out, the attitude of the federal courts to deal with the consequences of unrecognized or unapproached test item bias is itself a virtual enigma.

Given the current contravertible environment of test bias, its detection and correction, latent trait methodology--and specifically, the logistic response model--offers some appealing avenues for investigation. It is the intent of this study to further exploration of latent trait theory applications to test item bias methodology. The techniques used, their rationale and utility as applied to the current data set, is discussed. Detailed explanations of latent trait theory are described elsewhere (Hambleton, 1978; Lord, 1968; Warm, 1978; Wright, 1979a) and are not repeated here.

It is difficult to ignore the advantages latent trait theory offers over traditional psychometric methods in pursuing a fair, consistent, and workable definition and approach to the detection and correction of test item bias in widely used standardized tests. Of particular interest is the statistical independence of persons and test items. The separate estimation of these

parameters in the logistic response model approach (and its mathematical derivate) provides an avenue to avoid difficulties inherent in conventional biased item detection techniques; yet, still satisfied is the criterion of a consistent definition of item bias.

Scheuneman (1975) proposed a new operational definition which we shall adopt as containing sufficient rigor and accuracy for present purposes: "An item is considered unbiased if for persons with the same ability in the area being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual." This definition of bias is consistant with that used by Green and Draper (1972), and Pine and Weiss (1976). Scheuneman's definition describing the interaction of an examinee with a particular item provides a utilitarian way of detecting item bias in the context of, but not dependent upon, examinee performance.

The problem initially is one separating the parameters of persons and test items. The latent trait theory does this neatly and simply by proposing the model

$$\frac{P}{Q} = \frac{\alpha}{\xi}$$

where
- P = probability of a correct response
- Q = probability of an incorrect response
- α = person ability
- ξ = item difficulty

or, by log scale:

$$\ln\left(\frac{P}{Q}\right) = \ln\left(\frac{\alpha}{\xi}\right)$$

and, subtracting the difference of logs from this ratio yields:

$$\ln\left(\frac{P}{Q}\right) = \ln \alpha - \ln \xi$$

When a set of data is applied to the mathematical derivates of the model the statistics are easily calculated. The essential point for the present investigation is realized in their latently additive property. Hence, item difficulty can be separated from person ability. The methodology of item-free measurement continues on to describe precisely how person free item difficulties are estimated as well as item free person abilities. These calculations are described elsewhere (e.g., Rasch, 1961; An

are described elsewhere (e.g., Anderson, 1973, 1977: Baker, 1977: Hambleton, 1978: Rasch, 1961; Ryan, n.d.; Wright, 1977, 1979a, 1979b).

To the Rasch model, however, ln α and ln ξ are simply redefined as:

$$\beta = \ln \alpha, \text{ and}$$
$$\delta = \ln \xi$$

Hence, the derived equation, expressed probabilistically, is

$$P\{x_{vl}=1|\beta_v,\delta_l\}=\exp(\beta_v-\delta_l)/[1+\exp(\beta_v-\delta_l)]$$

where a person (v) with a defined ability (β_v) interacts to an item (l) with a calibrated difficulty (δ) to produce a response x_{vl}.

This is the only alternative to the models for a response curve which allows for independent estimations of person ability and item difficulty. Says Rasch: "When the estimators for β_v and δ_l are derived by maximizing a conditional likelihood they are unbiased, consistant, efficient and sufficient."

Item characteristic curves are computed by the regression of the test scores on ability β from a frequency distribution of test scores for each fixed level of β. Wright (1979) graphs the ogive for the theoretical response curve as :

Insert Table 1 here

Table 1

The Rasch Model Logistic Response Curve

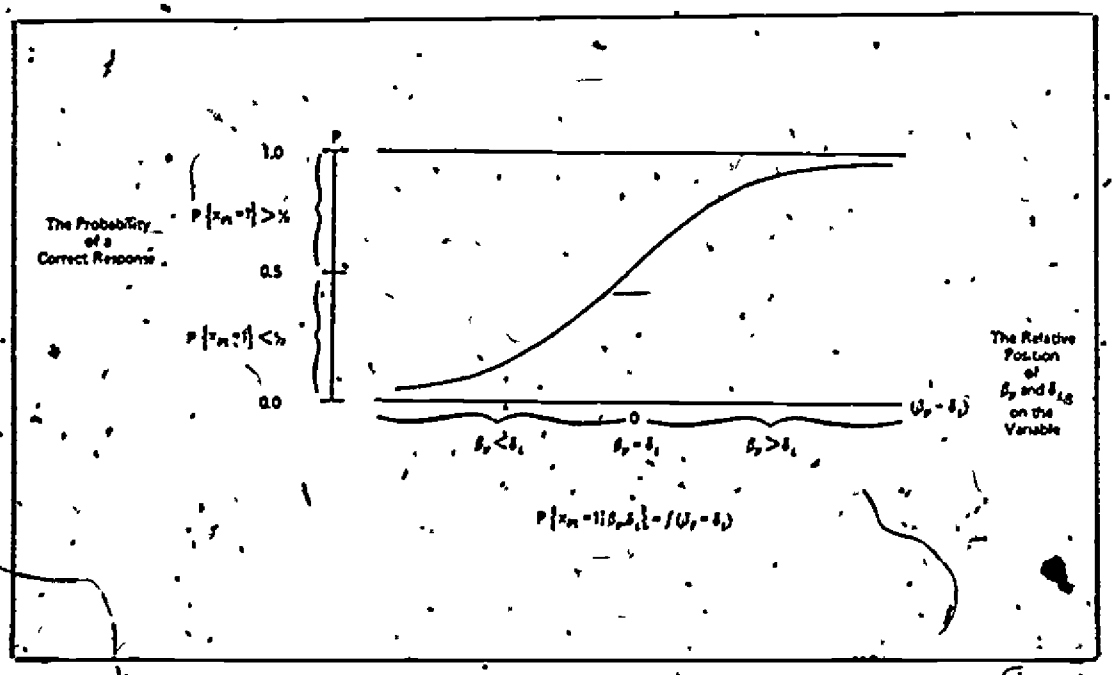


Figure from Best Test Design by B. D. Wright and M. H. Stone (MESA Press, 1979). Used with permission of the author.

With the logistic response model, at least three tests for validity are required. These are; 1) the more able a person the better the chance for success on any particular item; 2) any person has a better chance of correctly answering an easy item than a difficult one; and, 3) these conditions must be observably true regardless of any person's race, sex, or other noninterfering characteristic.

The third criterion is critical in test item bias methodology in that Scheuneman's definition of a biased test item is satisfied and it implies the notion of parameter independence. Herein lies the departure of the logistic response model approach to biased item detection from traditional techniques. The statistics conventionally employed in the search for a biased item are not independent of the sample ability distribution and are distorted for any independent analysis by this sample specific characteristic. Rudner (1980a) reviews several of the commonly used biased item detection techniques of empirical evidence of internal criteria in test bias; and, Peterson (n.d.) examines common arguments of bias in predictive validity. Each of the techniques and strategies discussed, however, is linked directly to the sample ability distribution. Tucker (1946) argues that this characteristic is not one that can enhance test rigor but actually may confound its own intentions. Wright (1976) demonstrates the point by citing a term--"sonata"--with high discrimination indices and is culturally skewed. Thus, the critical component of separate parameter estimation is not successfully addressed by any of the more traditionally used biased item detection statistics.

One further point is important to note for the present investigation. Latent trait theory and the logistic response model assumes local item independence. That is the performance of any examinee on any particular test item is an autonomous result of the interaction of pupil ability and item difficulty. The response by the examinee to that item is not influenced by a previous performance on any other item in the test. Lord (1953) demonstrated the validity of this assumption with a goodness-of-fit statistical test.

Item Bias Study

The present study of test item bias was conducted in a large west coast school district and utilized a single high school graduation proficiency examination. The test itself was developed with items selected from various recognized item banks along with a few new situation-specific items necessitated by the previously defined test content specifications. The items were then Rasch calibrated for goodness-of-fit of each to the model. Misfitting or ambiguously worded items were discarded. The calibrations were conducted with UCON estimation procedures.

It is to be noted that for the Rasch model calibration a single parameter assumption was made. Rudner (1977, 1980b) is critical of this assumption because ability in Rasch single parameter theory is based upon total score. Consequently, the presence of biased items aggregated into a total score could yield spurious results. He recommends adoption of the three-parameter model as developed by Birnbaum. This study did not accept the suggestion of a three-parameter model. Albeit conceded that a degree of rigor is added by the increased concern of an item discrimination index and a pseudo guessing parameter, the increased complexity as well as added difficulty of interpretation were not warranted in the present circumstance.

The high school graduation proficiency test was comprised of three subtests: reading comprehension, writing mechanics, and mathematics. A writing sample is also a required portion of the complete high school graduation proficiency test but it was scored by a holistic process and scores were not equated to Rasch scaling; and thus, it was excluded from the present study. For the subtests included, all questions were dichotomously scored multiple-choice questions. The reading test contained 30 items, the writing mechanics and mathematics tests had 35 each. Each subtest was treated independently of all others and item difficulty invariance was evaluated over each ability group for Black and White ethnic groups. Item plots for each ethnic group with a total group (i.e., all ethnic groups combined) were also examined.

The Sample Population

The sample population in the present application of latent trait theory to test item bias methodology included 5,309 reading comprehension tests, 5,284 tests of written expression, and 5,780 mathematics tests. By ethnic groups the distribution was less equal numerically although sufficient within each to yield valid results. The ethnic group populations were: 1,042 White, 11,441 Black, and 16,373 total group (including all Whites, all Blacks, and other unidentified). Table 2 presents these data as well as mean ability and standard deviation estimates for each ethnic group by subtest.

Insert Table 2 about here

Table 2

Ethnic Group by Subtest:

Number, Mean Ability, and Standard Deviation Ability

Test	Ethnic Source Group					
	White		Black		Total*	
	N	Ability: $\bar{X}$ and σ	N	Ability: $\bar{X}$ and σ	N	Ability: $\bar{X}$ and σ
Reading (30 items)	308	2.19 .86	3717	1.14 .89	5309	1.21 .96
Written Expression (35 items)	384	1.89 1.05	3557	.68 .92	5284	.83 1.03
Mathematics (35 items)	350	2.45 1.04	4167	1.25 1.01	5780	1.40 1.12
TOTAL	1042		11,441		1,6373	

*Includes total White, total Black, and other unidentified

Item Difficulty Estimations

For the best calibration persons should be about evenly distributed over the range of scores around and above the center of the test (Wright, 1979a). The sample person ability distribution data were computed; and, as the data reveal, scores are not symetrically distributed but negatively skewed around and above a modol raw score of 22 to 25 in reading comprehension, 19 to 26 in written expression, and 29 to 30 in mathematics. This result was anticipated due to the nature of item content specifications for minimal high school graduation skills rather than allowing for a range of abilities from quite low to very high. (The frequency distribution tables for each subtest are included in Appendix B.)

Item Plots

The constructed item plots for each of the items on every subtest allow inspection of the extent to which the item points conform to the model expectation of item difficulty invariance. This inspection of item invariance across different ethnic groups is a measure of the quality of individual items to be free from or contaminated by some degree of bias.

Each pair of calibrations applies to one, and only one, item, and of course, two difficulties (d_{i1} and d_{i2}) were derived. Standard errors for each item (s_{i1} and s_{i2}) was also computed. Hence, only a single translation is necessary to establish an origin common to both sets of items at any difficulty, δ_i.

Wright (1979a) gives the statistic for testing the estimate of δ_i by d_{i1} and d_{i2}. It is:

$$t_{i12} = (d_{i1} - d_{i2})/(s_{i1}^2 + s_{i2}^2)^{\frac{1}{2}} \sim N(0,1)$$

Tests for the quality of fit of each item (viz., item invariance across ethnic group population calibrations) can be made by positioning quality control boundaries at about two standard errors away from an identity line on each side. Two of these quality control boundaries parallel to the initial identity control line approximate a 95% confidence boundary. This is calculated by the formula

$$D_{i12} = 2[(s_{i1}^2 + s_{i2}^2)/2]^{\frac{1}{2}}$$

$$= 2[(s_{i1} + s_{i2})/2] = s_{i1} + s_{i2}.$$

D_{i12} is the perpendicular distance between the quality control line and the identity line. The formula $(s_{i1}^2 + s_{i2}^2)^{\frac{1}{2}}$ estimates the standard error of the difference between the two independent estimates d_{i1} and half of this, or $[(s_{i1}^2 + s_{i2}^2)/2]^{\frac{1}{2}}$, is the error unit perpendicular to a 45 degree identity line.

These control lines for identity plots may be graphically presented as follows.

Insert Table 3 about here

Table 3

Estimation of Quality Control Lines.

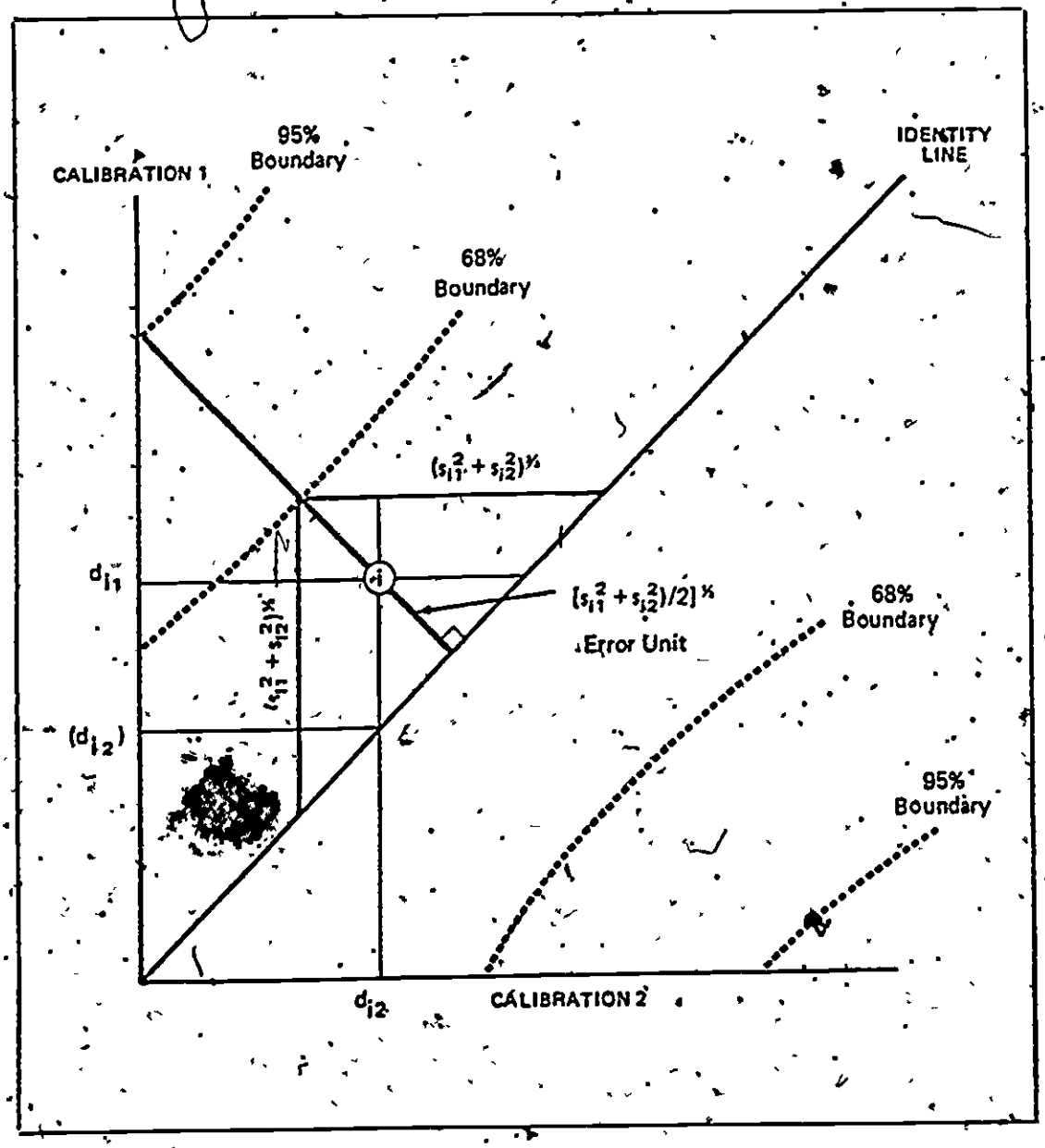


Figure from Bebs Test Design by B. D. Wright and M. H. Stone (MESA Press, 1979). Used with permission of the author.

Table 4 presents the sample population groups by subtest upon which fit statistics were computed. The calibrations were UCON estimations. (Technically, last difference change and comparisons with PROX procedures were also calculated. The tables for each ethnic group by subtest are included in Appendix A.) Table 5 displays the schema of study design in which the nine item plots were constructed.

Insert Table 4 and Table 5 about here

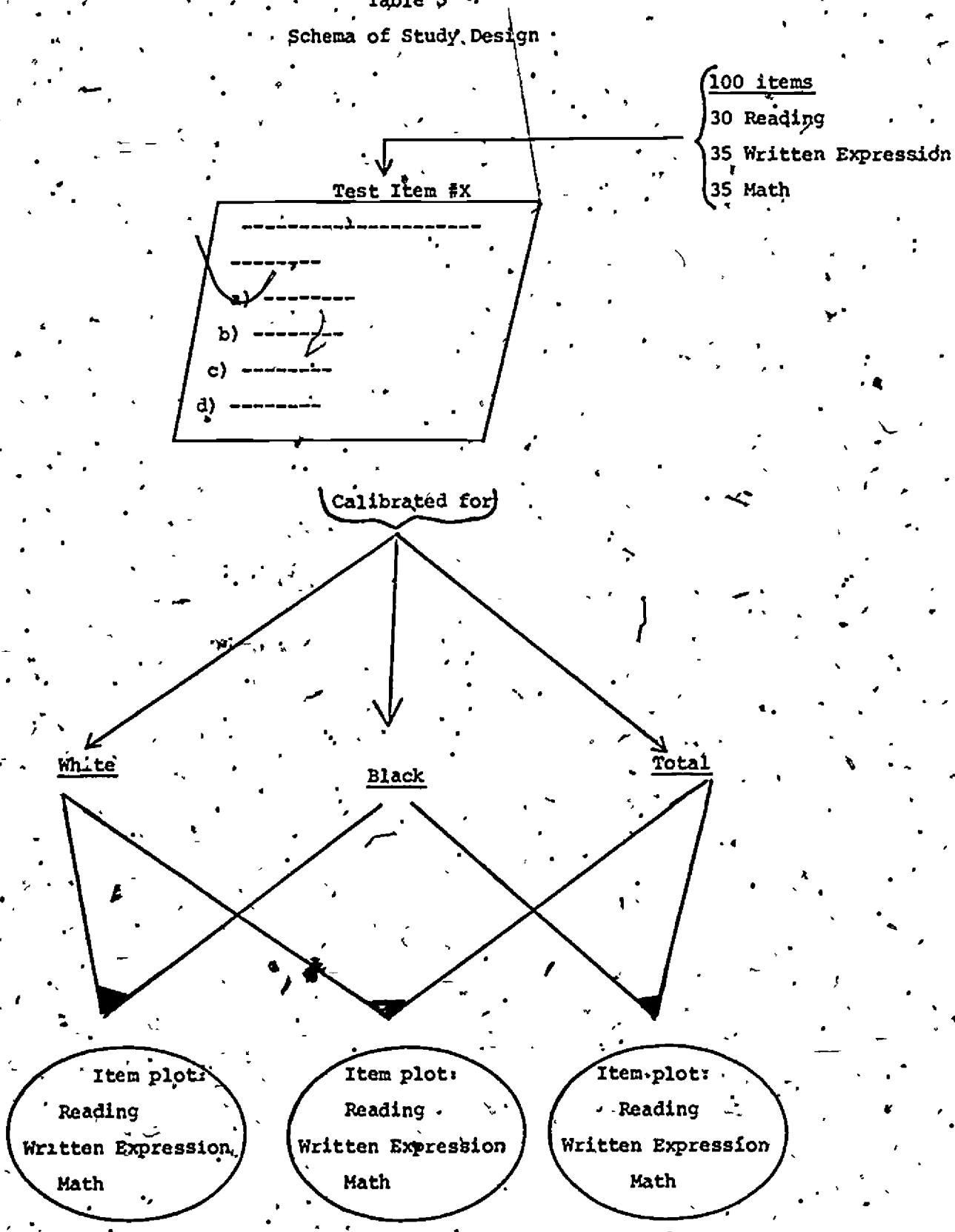
Table 4

Sample Population Group by Subtest

No.	Source	Subtest
1	White x Black	Reading
2	White x Total	Reading
3	Black x Total	Reading
4	White x Black	Written Expression
5	White x Total	Written Expression
6	Black x Total	Written Expression
7	White x Black	Mathematics
8	White x Total	Mathematics
9	Black x Total	Mathematics

Table 5

Schema of Study Design



Data Analysis

An examination of each of the nine item plots reveals a remarkable degree of item invariance for each of the items on all subtests. Two of the item plots are displayed: Table 6 presents the subtest exhibiting the least item invariance (ethnic group Black versus ethnic group Total: Math) and Table 7 displays the item plot wherein item invariance beyond confidence limits is revealed (ethnic group White versus ethnic group Black. Written Expression). (The remaining item plots are included in Appendix C.)

- -

Insert Table 6 and Table 7 about here

- -

MATH
Black x Overall

TABLE 6

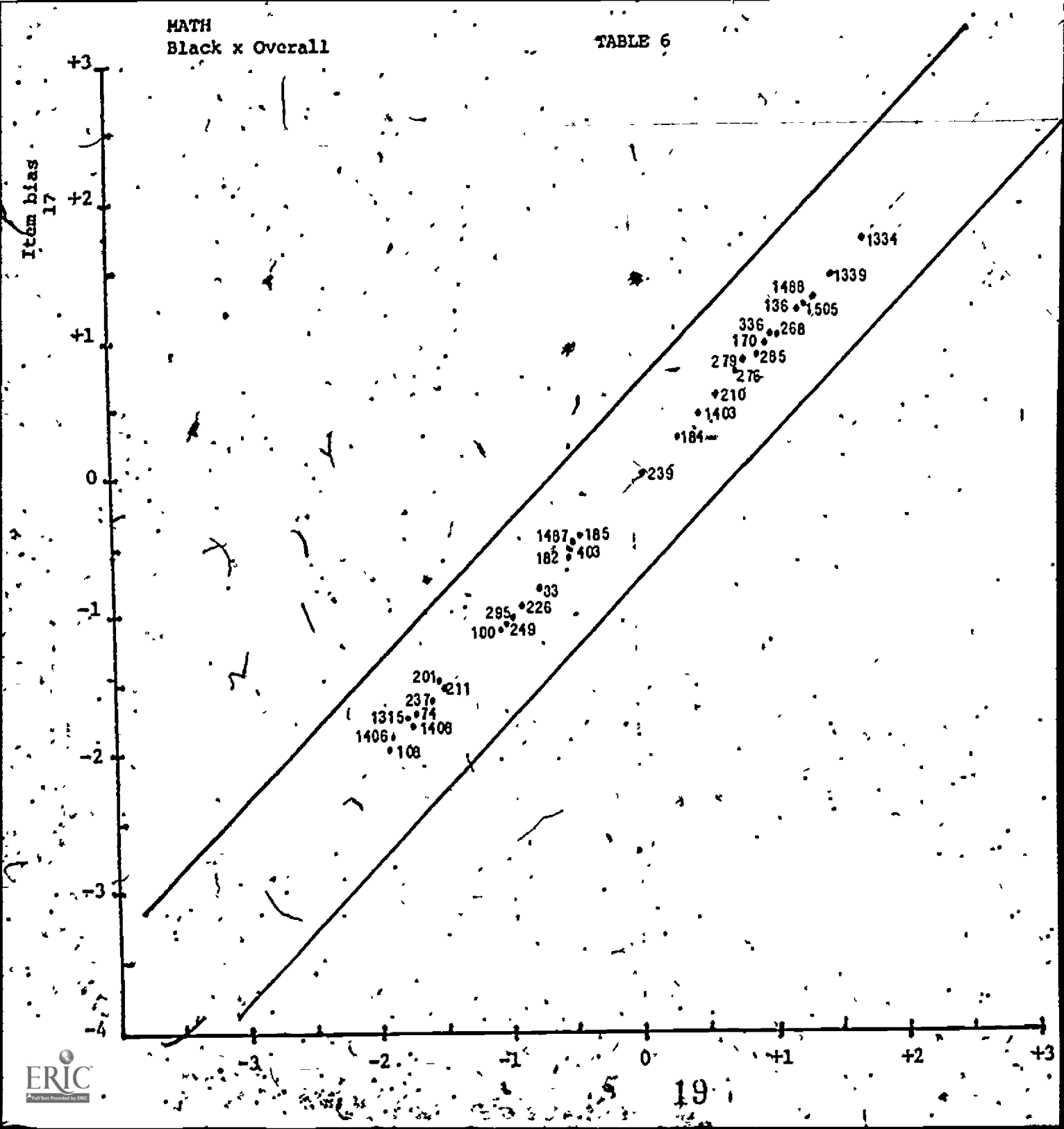
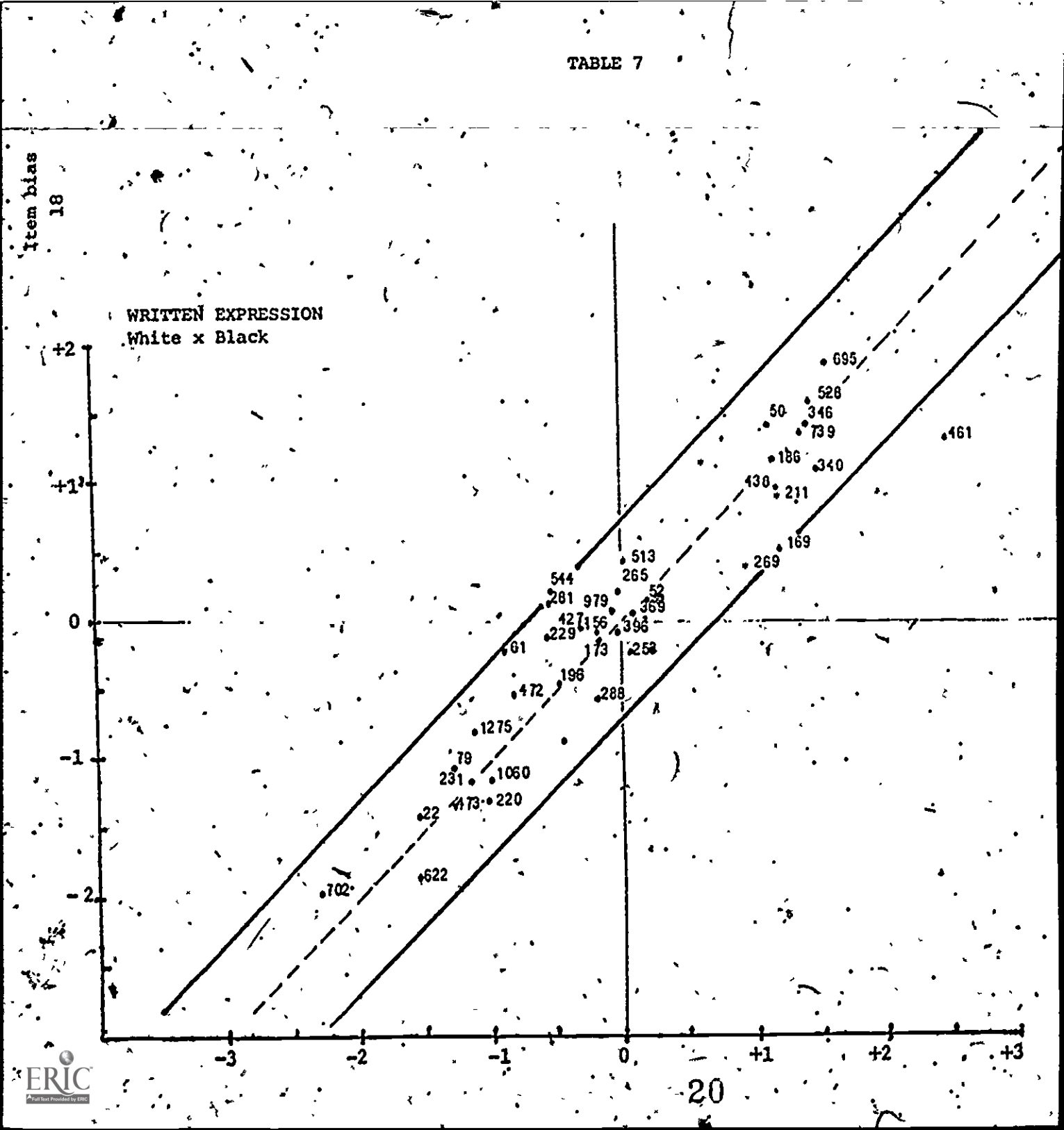


TABLE 7


Item bias
18
WRITTEN EXPRESSION
White x Black
+2
+1
0
-1
-2
-3
-2
-1
0
+1
+2
+3
695
526
50
346
739
461
186
340
430
211
169
269
513
544
265
281
979
52
369
427
156
396
229
61
173
253
196
472
288
1275
79
231
1060
220
22
173
622
702

The reading subtest contained least item invariance for each ethnic group comparison. The mathematics subtest also exhibited minimal item invariance despite the largest item difficulty range among the three subtests (-2.339 to 4.296 logits).

Within the writing subtest, however, a single item did exhibit an unacceptable degree of item invariance. This item plotted for ethnic group White versus ethnic group Black outside of quality control lines, and thus represented the detection of an item confounded by ethnic group calibration. The item, identified as Item No. 461, calibrated at 1.327 logits difficulty and standard error of .110 for ethnic group White and nearly twice as large at 2.50 logits difficulty and standard error of .048 for ethnic group Black.

The large difference in item difficulty estimates between calibration by ethnic group White and ethnic group Black and the resultant outlier characteristic on the item plot pointed to an inspection of the item wording. The item read as follows:

Insert Figure 1 about here

Figure 1

Item No. 461

Select the word or groups of words that correctly completes the sentence.

The grasshoppers in our garden _____ the vegetables.

A is eating
B. eats
*C eat
D does eat

Curriculum experts examined the items and surmised that the correct response C *(eat)* may be confounded by modest dialectical differences among ethnic group Black examinees. Traditional item statistics support this supposition. As revealed by p- values, ethnic group Black examinees missed the item much more often than did ethnic group White examinees, and response B *(eats)* was the most frequently selected distractor by ethnic group Black. Not surprizingly, analysis of variance revealed that while between group variance was large, within group variance was very small. Yet, in total group the item held a high discrimination index (point biserial). Thus, in this study of item bias detection, it is likely that this particular item may have been overlooked in a search for biased items using traditional statistics; yet, with the logistic response model of latent trait theory, this defective item was detected and removed from the test.

References

Andersen, E. B. A goodness of fit test for the Rasch Model. Psychometrika, 1973, 38, 123-140.

Baker, F. B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Fincher, C. Differential validity and test bias. Personnel Psychology, 1975, 28, 481-500.

Green, D. R. and Draper, J. F. Exploratory studies of bias in achievement tests. Paper presented at the American Psychological Association Annual Convention, Honolulu, Hawaii, September, 1972.

Hambleton, R. K. et al. Developments in latent trait theory: models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Jensen, A. R. Bias in mental testing. New York: The Free Press, 1980

Lord, F. M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548. (Also Research Bulletin 52-10. Princeton, NJ.: Educational Testing Service, 1952.)

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813.

Pine, S. M. and Weiss, D. J. Effects of item characteristics on test fairness, Research Report 76-5, Minneapolis: University of Minnesota Psychometric Methods Program, December, 1976.

Rasch, G. On general laws and the meaning of measurement in psychology. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1961, 4, 321-333.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the Annual Meeting of The American Educational Research Association, New York, April, 1979.

Rudner, L. M., Getson, P. R., and Knight, D. L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233. (a)

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, 17, 1-10. (b)

Ryan, J. P. and Hamm, D. W. An introduction to the Rasch latent trait psychometric model, College of Education, University of South Carolina n.d.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at American Educational Research Association, Washington, DC., 1975.

Scheuneman, J. A new method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-14.

Warm, T. A. A primer of item response theory. U. S. Coast Guard Institute. Oklahoma City, OK, December, 1978

Wright, B. D., Mead, R. and Draba, R. Detecting and correcting test item bias with a logistic response model, Research Memorandum Number 22, October 1976, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. and Douglas, G. W. Conditional versus unconditional procedures for sample-free item calibrations. Educational and Psychological Measurement, 1977, 37, 47-60.

Wright, B. D. and Stone, M. H. Best test design. Chicago: MESA press, 1979. (a)

Wright, B. D., Mead, R. J., and Bell, S. R. BICAL: Calibrating items with the Rasch Model. Research Memorandum Number 23B, June 1979, Statistics Laboratory, Department of Education, University of Chicago. (b)

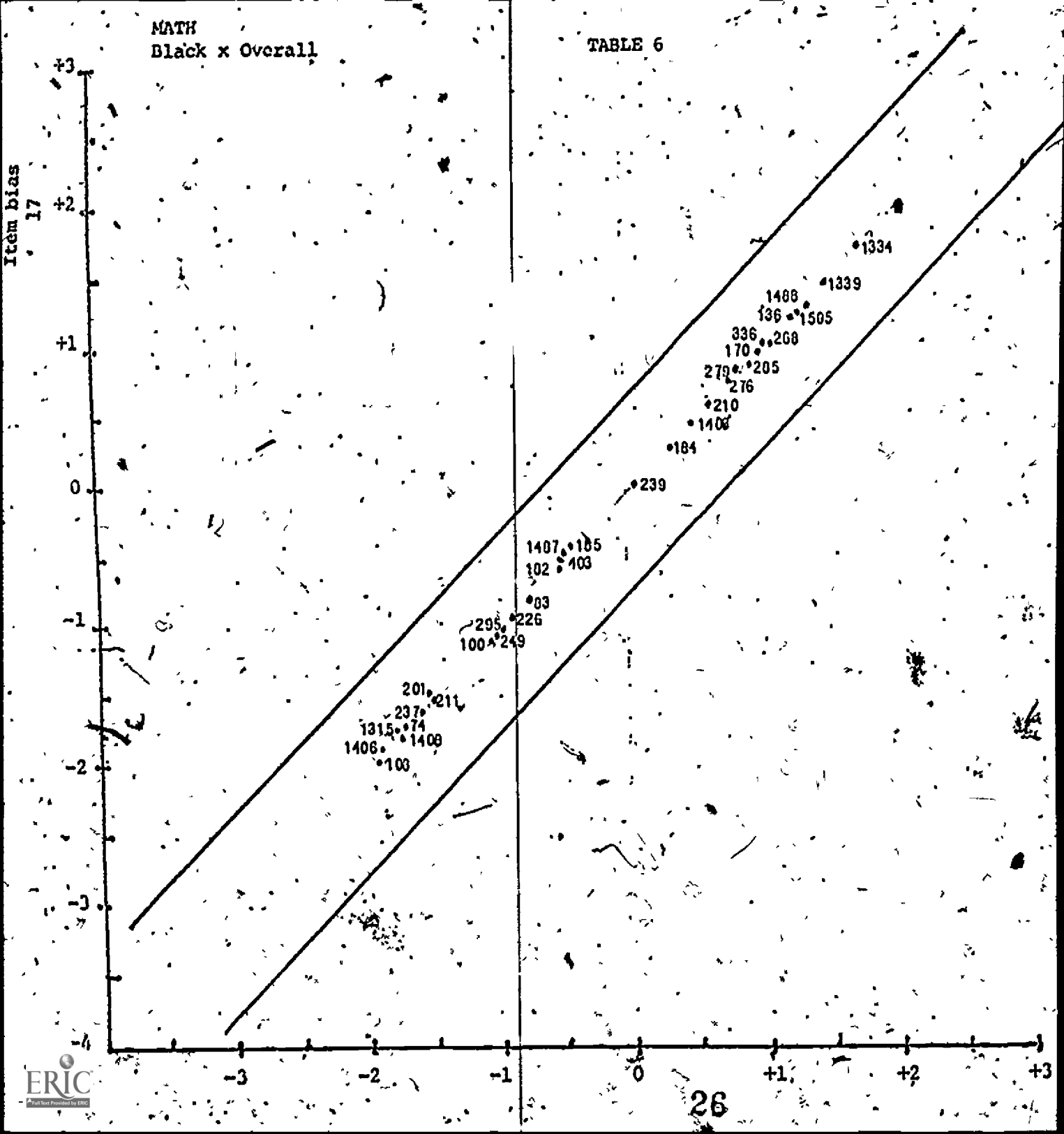
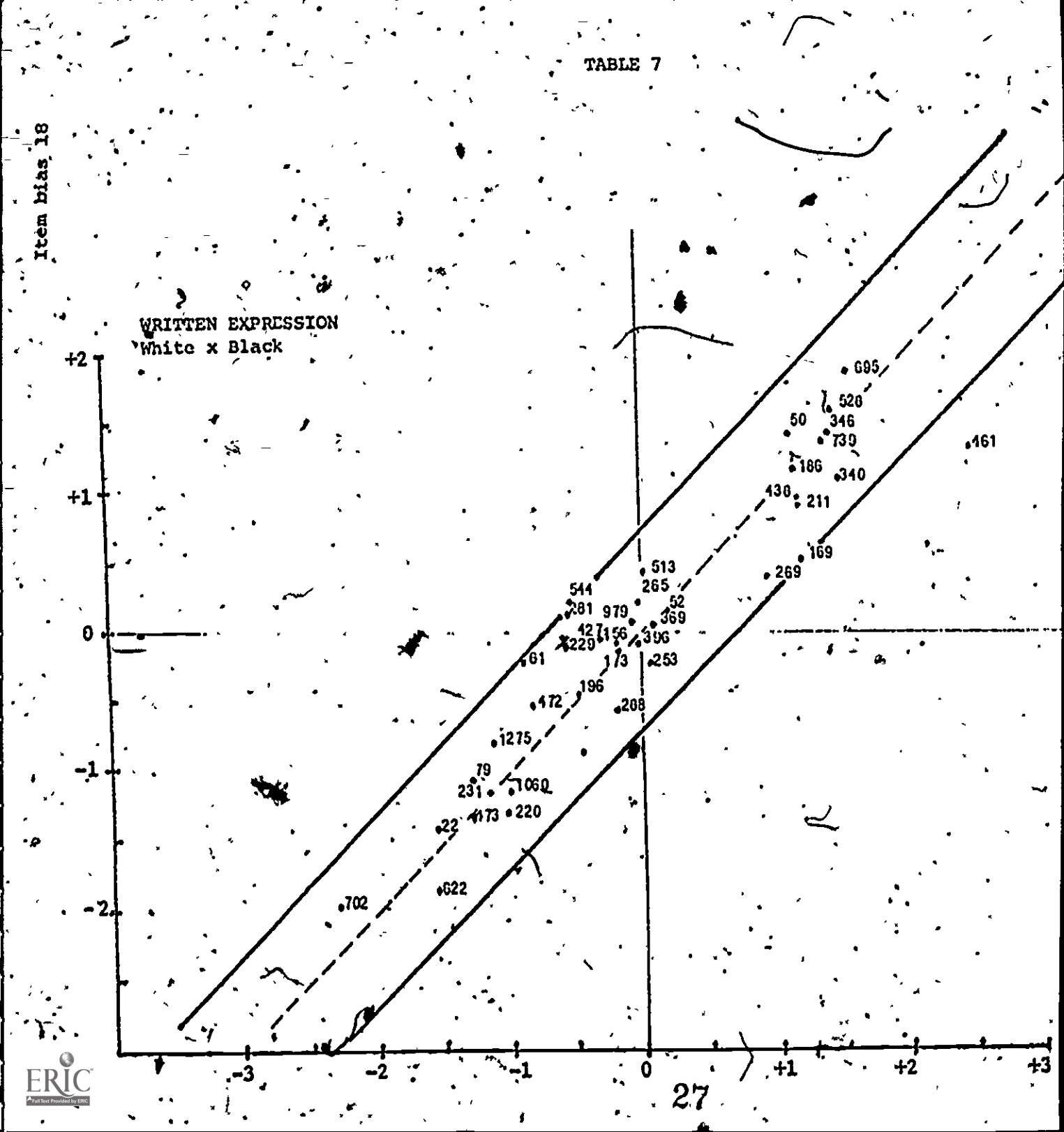
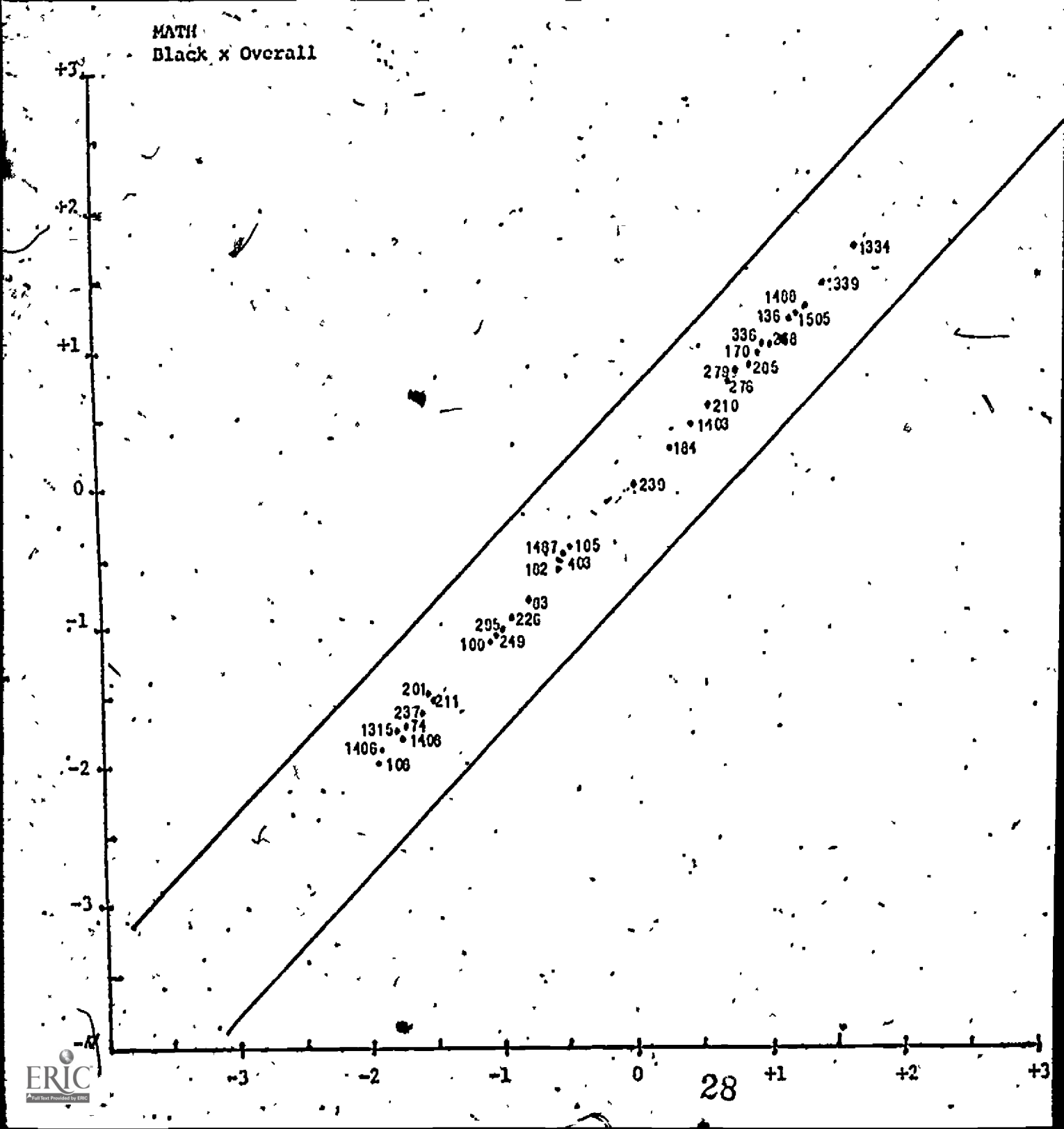
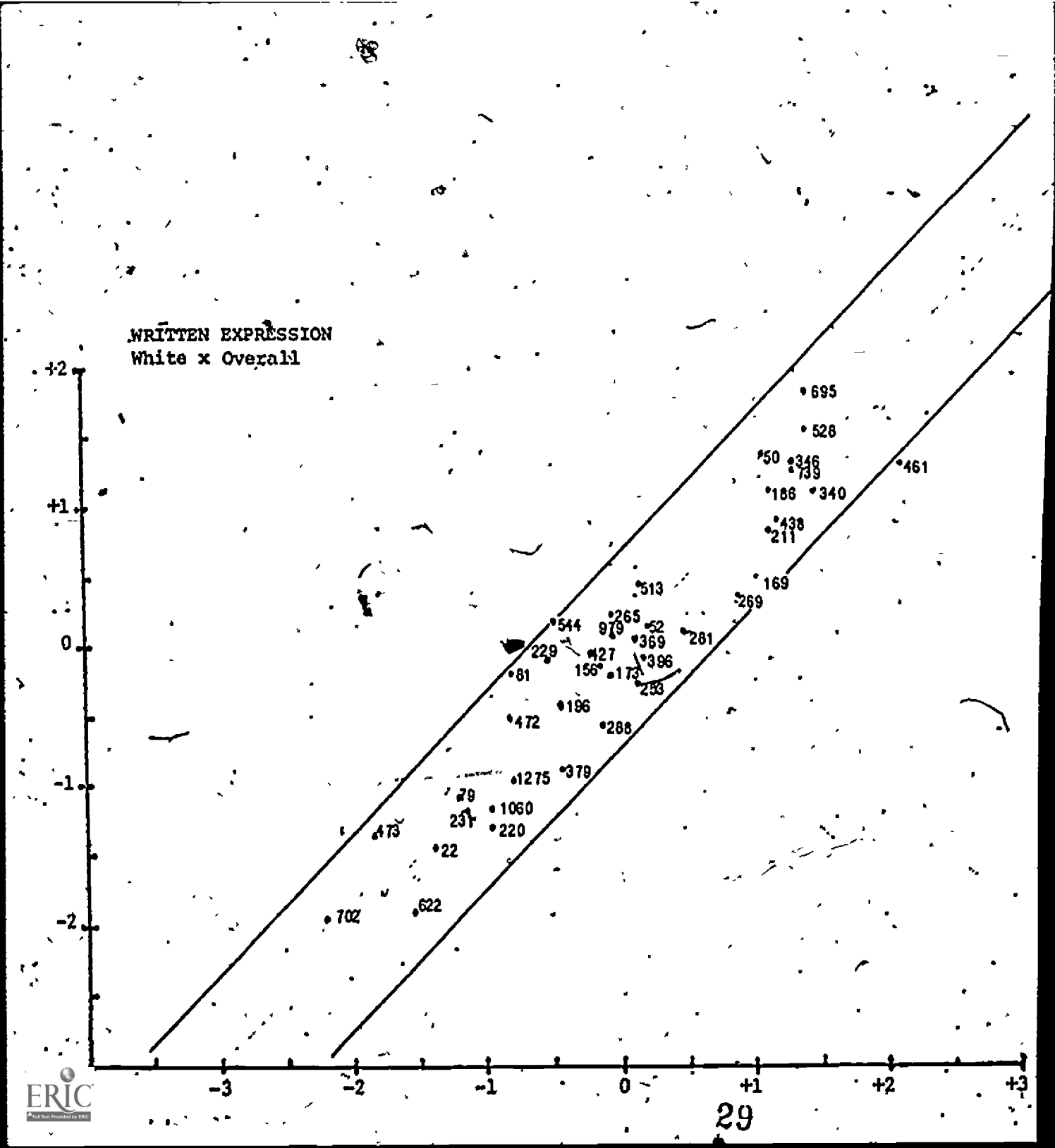
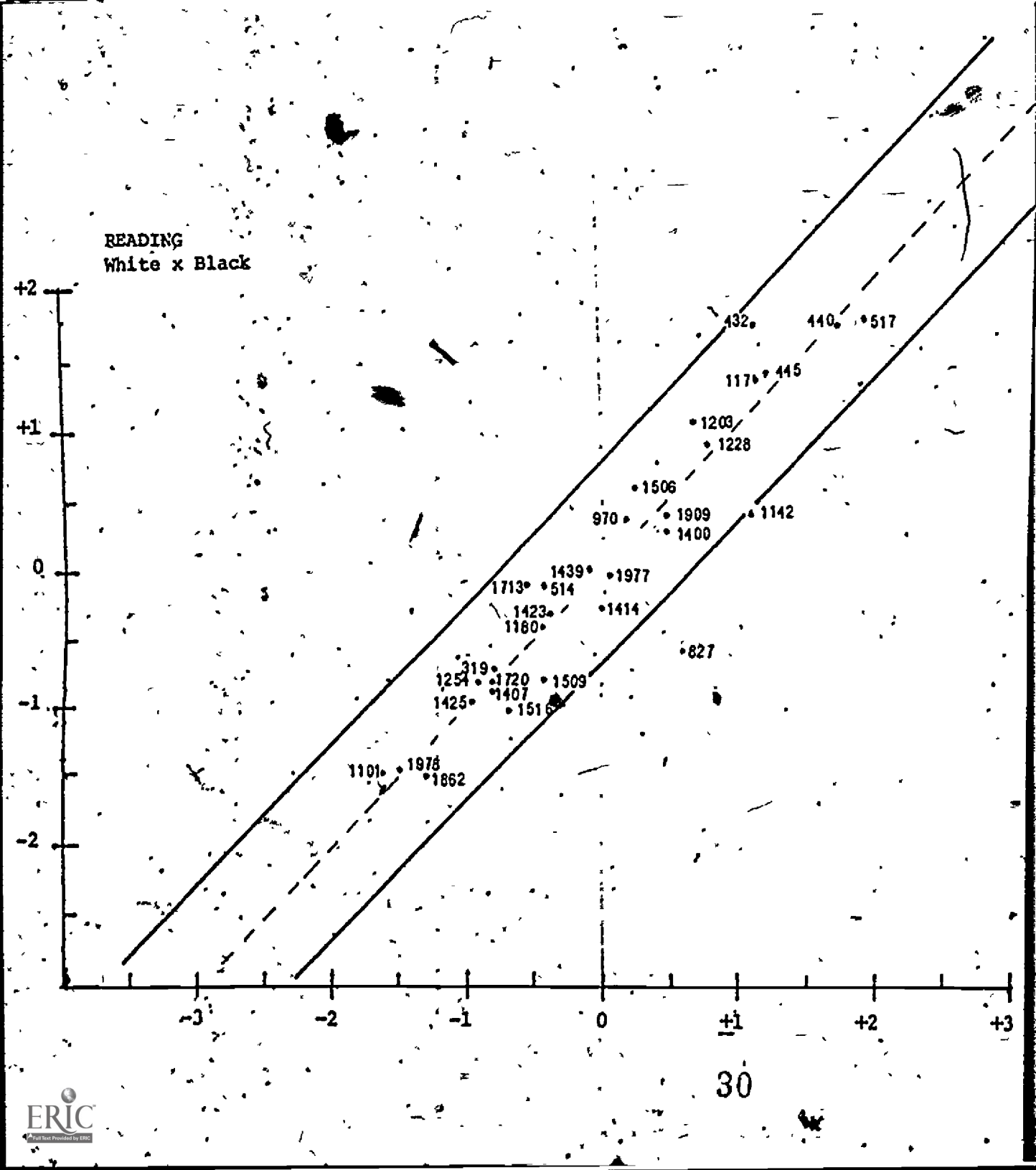

MATH
Black x Overall
TABLE 6
Item bias
17
+3
+2
+1
0
-1
-2
-3
-4
-3
-2
-1
0
+1
+2
+3
1334
1339
1488
136
1505
336
208
170
205
279
276
210
1408
184
239
1407
185
102
103
83
226
295
100
249
201
211
237
74
1315
1409
1406
103

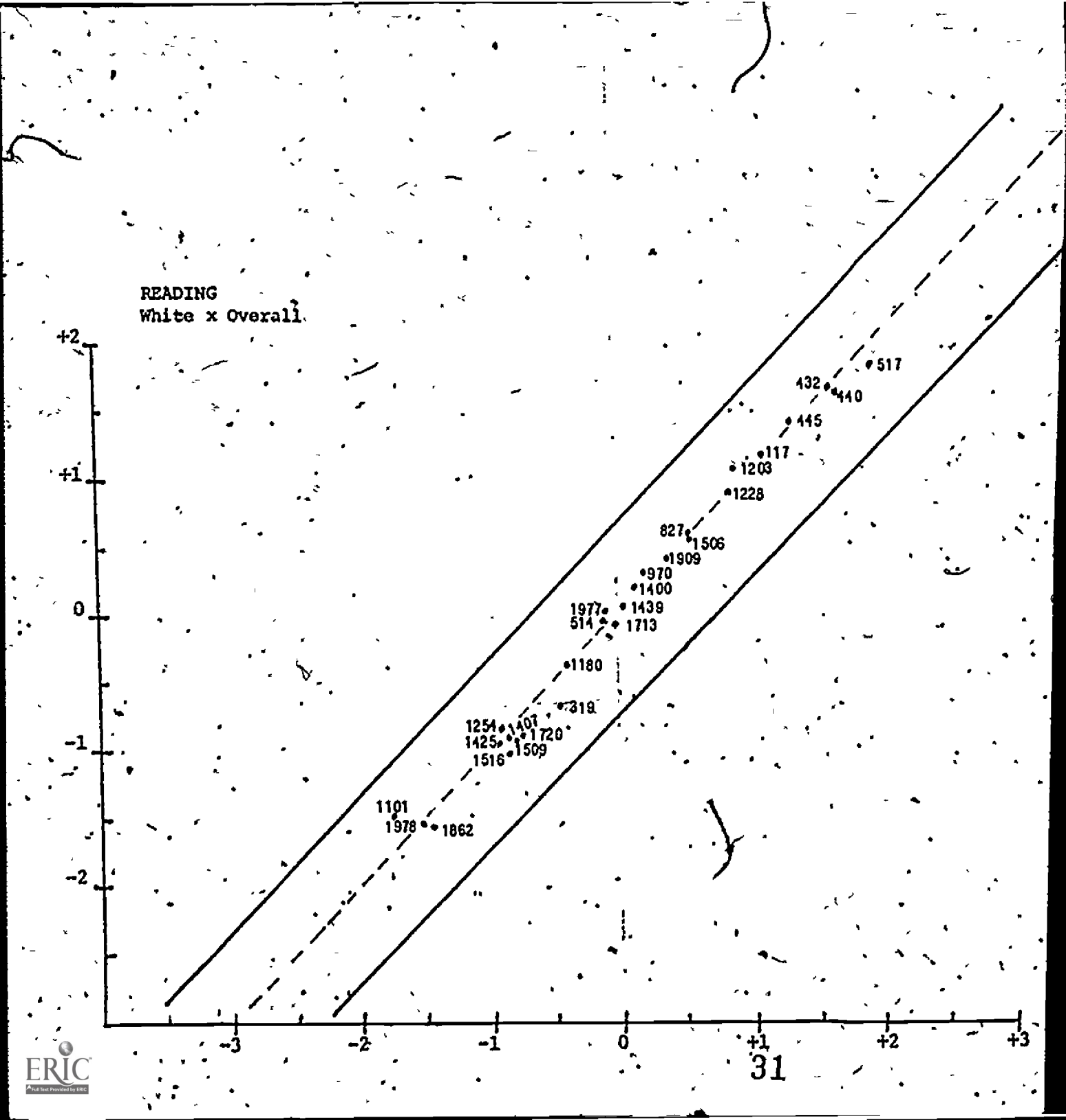
TABLE 7

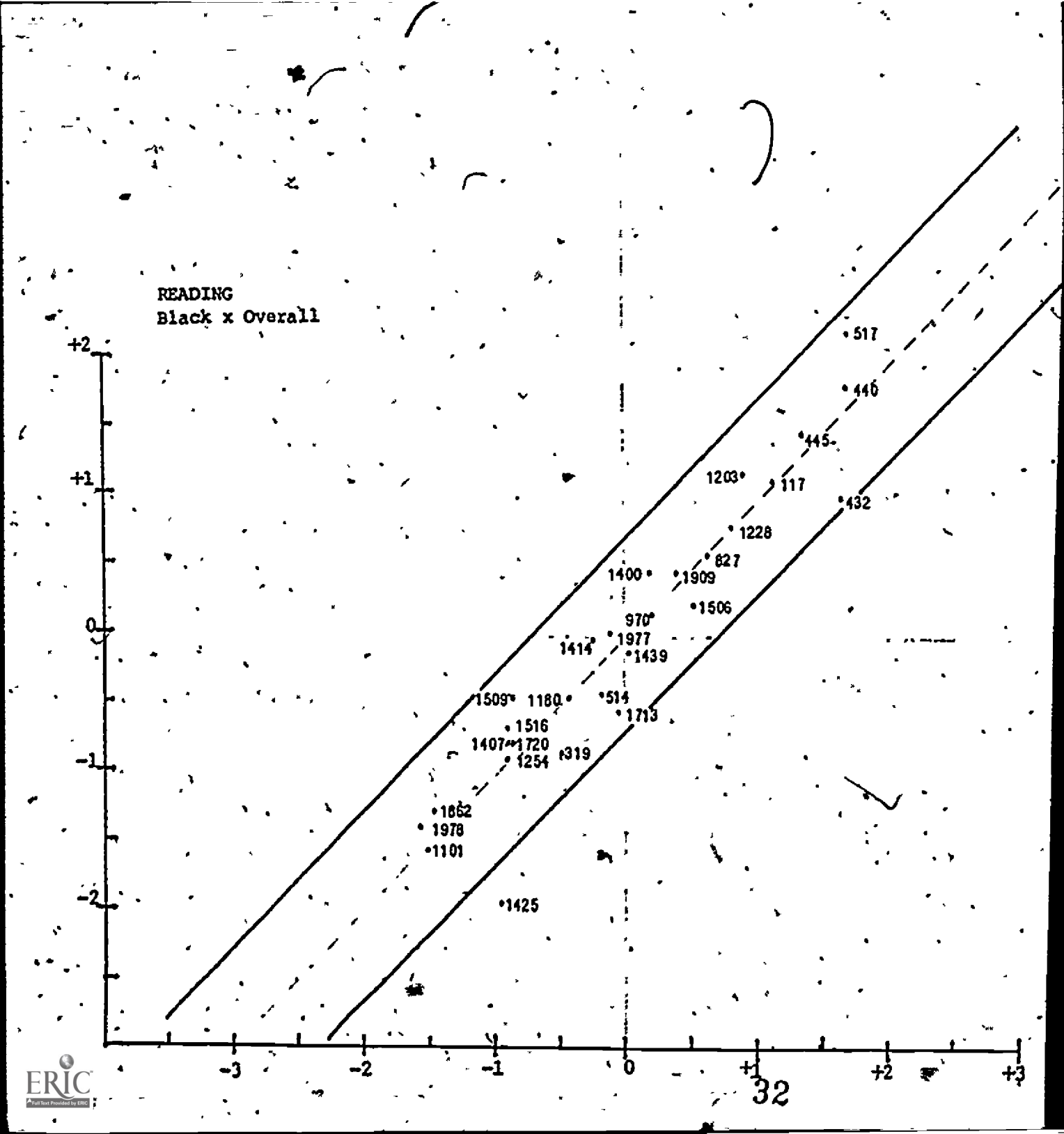

Item bias 18
WRITTEN EXPRESSION
White x Black
+2
+1
0
-1
-2
-3
-2
-1
0
+1
+2
+3
695
520
50
346
739
461
186
340
436
211
169
269
513
265
544
52
281
979
369
427
156
396
229
61
173
253
196
472
208
1275
79
231
1060
173
220
22
702
622

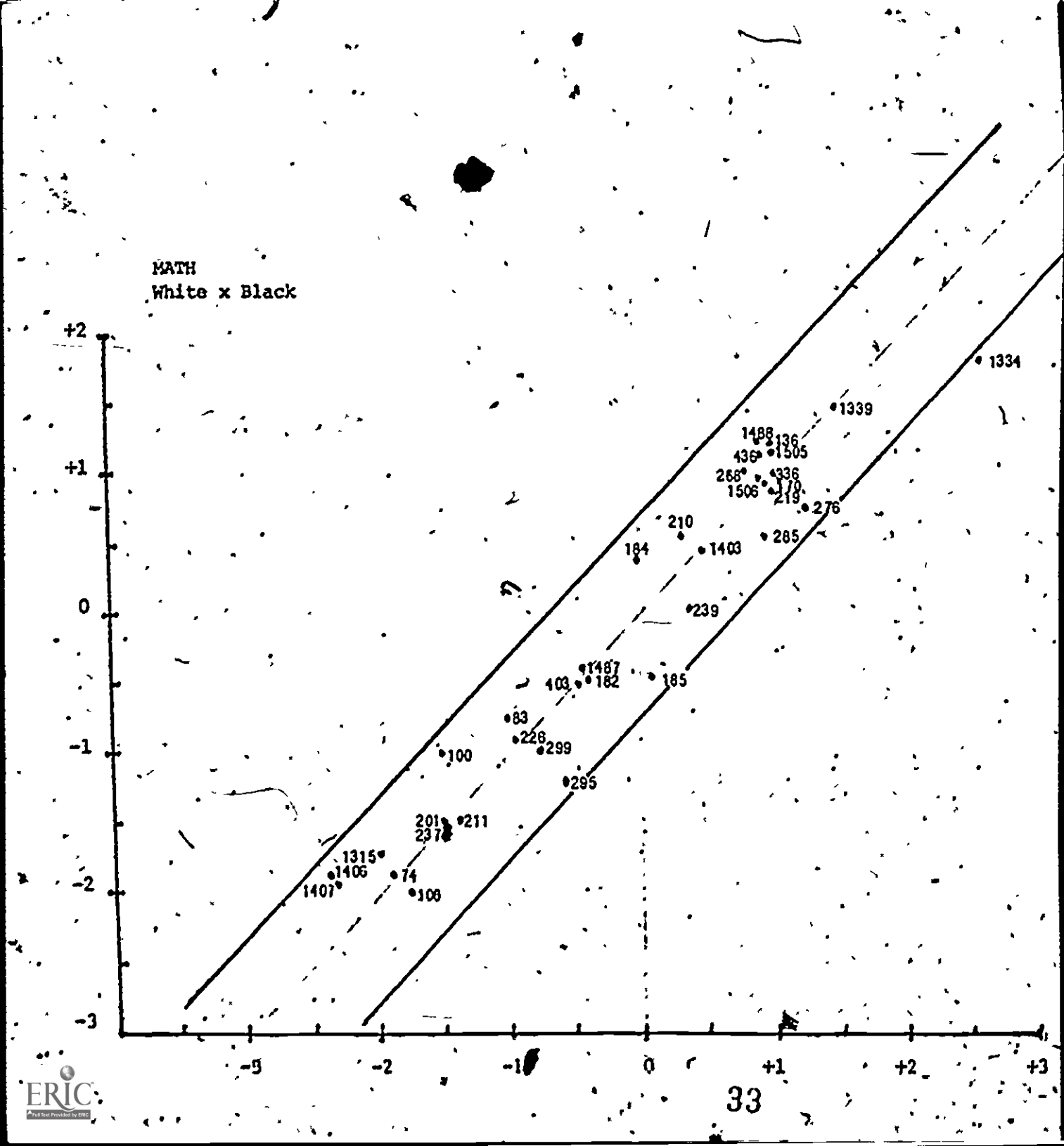

MATH
Black x Overall
+3
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
1334
1339
1408
136
1505
336
268
170
205
279
276
210
1403
184
239
1407
105
102
403
03
226
295
100
249
201
211
237
74
1315
1406
1406
108

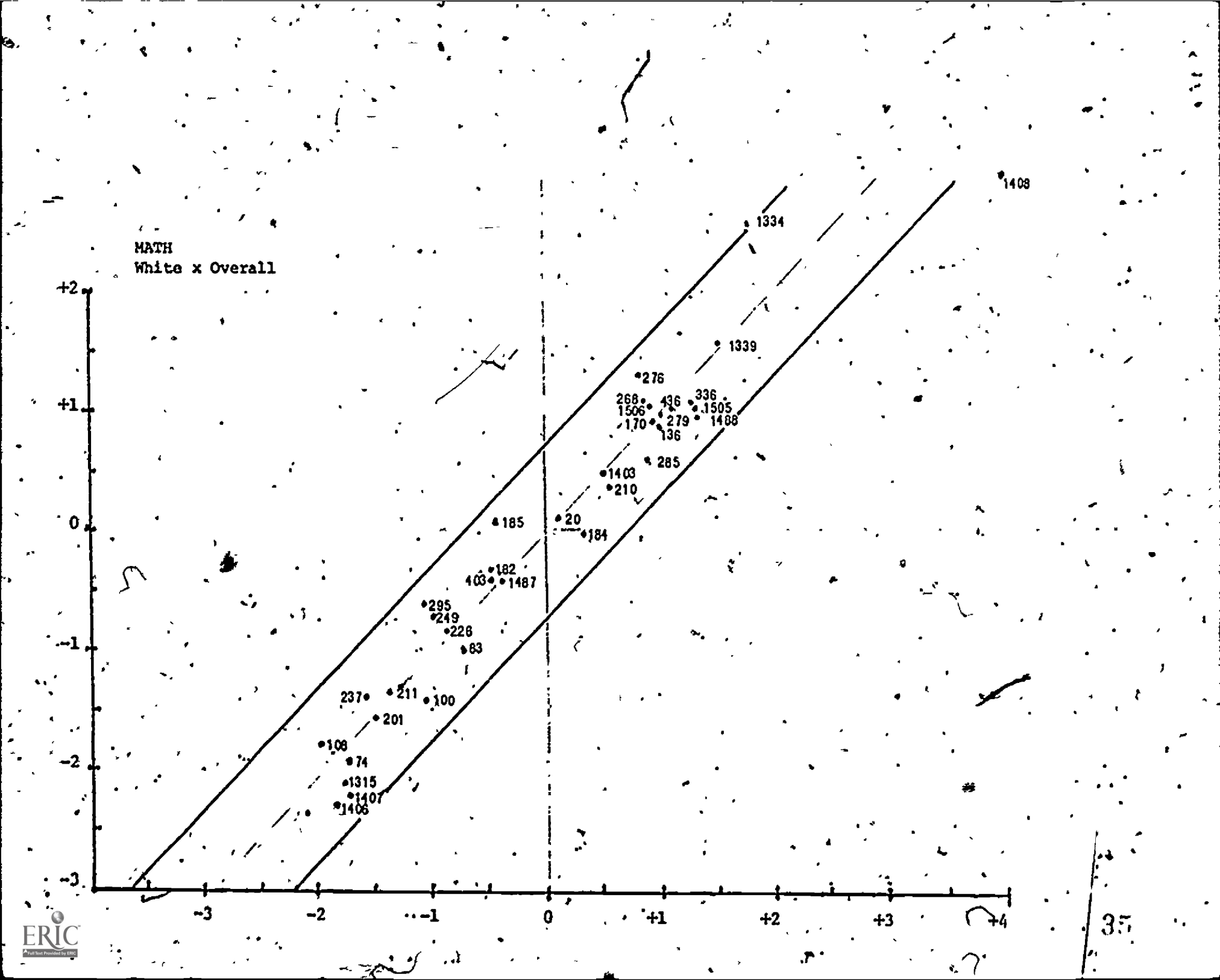

WRITTEN EXPRESSION
White x Overall
+2
+1
0
-1
-2
-3
-2
-1
0
+1
+2
+3
695
528
50
346
739
461
186
340
438
211
513
169
269
544
979
265
52
369
261
229
427
396
81
155
173
253
196
472
288
1275
379
79
1060
231
220
473
22
702
622


READING
White x Black
+2
+1
0
-1
-2
-3
-2
-1
0
+1
+2
+3
432
440
517
445
1203
1228
1506
970
1909
1400
1142
1439
1977
1713
514
1423
1414
1180
827
319
1254
1720
1509
1407
1425
1516
1101
1978
1862


READING
White x Overall
+2
+1
0
-1
-2
-3
-2
-1
0
+1
+2
+3
517
432
440
445
117
1203
1228
827
1506
1909
970
1400
1977
1439
514
1713
1180
319
1254
1407
1425
1720
1509
1516
1101
1978
1862


READING
Black x Overall
+2
+1
0
-1
-2
-3
-2
-1
0
+1
+2
+3
517
446
445
1203
117
432
1228
827
1400
1909
1506
970
1977
1414
1439
1509
1180
514
1713
1516
1407
1720
319
1254
1652
1978
1101
1425


MATH
White x Black
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
1334
1339
1488
136
1505
436
266
336
1506
170
219
276
210
285
184
1403
239
1487
403
182
185
83
226
299
100
295
201
211
237
1315
1406
74
1407
100


MATH
White x Overall
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
+4
1408
1334
1339
276
268
436
336
1506
1505
170
279
1488
136
285
1403
210
185
20
184
182
403
1487
295
249
226
83
237
211
100
201
108
74
1315
1407
1406