DOCUMENT RESUME

ED 323 213 TM 015 359

AUTHOR Sykes, Robert C.; Fitzpatrick, Anne R.

TITLE Establishing a Mantel-Haenszel Alpha Cutscore through a Multiple Method Procedure.

PUB DATE Apr 90

NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Classification; *Cutting Scores; Difficulty Level; Ethnic Groups; *Ethnicity; Evaluation Methods; *Item Response Theory; Licensing Examinations (Professions); Scores; *Test Items

IDENTIFIERS *Alpha Estimates; *Mantel Haenszel Procedure; Standard Setting

ABSTRACT

The results of classifying test items on the basis of their Mantel-Haenszel (MH) alpha estimates were compared to the results of classifying these items using an item response theory (IRT) based procedure involving the comparison of item difficulties in the interest of identifying the alpha value that maximized the decision concordance between the two methods. The data consisted of candidates' responses to 299 scored items on an examination for professional licensure. A total of 68,458 candidates took this examination in 1988. The candidates' ethnicity was determined by their self-classifications into one of seven categories. A total of 47,573 candidates classified themselves in Ethnic Group 1; 6,486 in Ethnic Group 2; 5,466 in Ethnic Group 3; 2,004 in Ethnic Group 4; 1,014 in Ethnic Group 5; 486 in Ethnic Group 6; 307 in Ethnic Group 7; 726 as "other"; and the remaining 4,396 candidates did not specify their ethnicity. All candidates except the 4,396 who did not classify themselves were used for the MH analyses; for the IRT analyses, random samples of candidates were drawn from the four largest ethnic groups, while all candidates from the smaller groups were used. Candidates were divided into score groups. In both analyses, the majority ethnic group became the reference group and the other ethnic groups were designated focal. The MH estimate of alpha indicated that no differential item functioning was detected. Statistics resulting from IRT methods correlated highly with MH alpha values. Using the MH method in conjunction with IRT methods resulted in cutscores with a high level of decision concordance with advantages over traditional methods of using a significance level to establish an alpha criterion. Eight tables and eight figures provide study data. (SLD)

Establishing a Mantel-Haenszel Alpha Cutscore
Through a Multiple Method Procedure

Robert C. Sykes
Anne R. Fitzpatrick

CTB/Macmillan/McGraw-Hill School Publishing Co.

This paper was presented in April, 1990 at the annual meeting of the American Educational Research Association, Boston, MA.

The problem of assessing bias in test items has concerned measurement specialists for many years, and they have proposed a variety of methods for detecting items that function differentially for majority and minority groups. Berk (1982), Cole and Moss (1989), and Hills (1989) have provided some of the most recent reviews of these methods.

Among the methods available, the Mantel-Haenszel (MH) method has recently received a great deal of interest as a practical means of assessing uniform bias (Hambleton & Rogers, 1989; Hills, 1989). As described by Holland and Thayer (1986) this method involves computing for the majority group and a minority group (called the reference group and focal group, respectively) the odds ratio (called alpha) of their success on an item across score groups in which the group members' ability levels are held constant. If an item is not functioning differentially, the reference group and focal group members having the same ability will perform equally well on the item, in which case alpha (α) will be equal to 1.0. To the degree that the performance of the two groups differs, alpha will deviate from 1.0, and the item can be said to exhibit differential item functioning (DIF).

An important question for users of the MH method who must analyze large samples is what constitutes a meaningful amount of DIF. The significance level associated with a chi-square test of the null hypothesis that $\alpha = 1.0$ can be used to distinguish statistically significant levels of alpha (see Holland & Thayer, 1986), but this significance test is sensitive to the effects of sample size, as significance tests generally are. More specifically, if large enough samples are used, the null hypothesis can be rejected even when the differences in the item performance of the majority group and the minority group are of little practical importance. Also, the use of a significance test for detecting items that show DIF will result in more items detected when large samples are analyzed than will be detected when small samples demonstrating the same amount of "true" DIF are analyzed.

An alternative to using the significance level associated with a chi-square test for classifying items is using the standard errors of the alpha estimates to define what is "meaningful" DIF (Phillips & Holland, 1987). However the standard error is also sensitive to sample size; it decreases as sample size increases. Consequently the number of items classified as having DIF again will vary with the sizes of the samples being analyzed. Also this number will vary with the number of standard errors that is somewhat arbitrarily selected to construct a confidence interval that can be used to distinguish meaningful levels of alpha from non-meaningful levels.

A third approach to identifying a meaningful level of DIF is used in the current study. This approach involves the use of multiple methods for measuring DIF in the interest of identifying that level of DIF at which these methods produce consistent classifications. This approach has its basis in the logic of the multi-trait multi-method approach first described by Campbell and Fiske (1959). According to this logic, a test that is purported to measure a given construct should show a strong relationship to other measures of the same construct (Messick, 1989). The strength of the relationship between the test and these other measures comprises evidence of the degree to which the construct exists independently of the method used to measure it. As explained by Denzin (1978),

> The rationale for this strategy is that the flaws of one method are often the strength of another, and, by combining methods, observers can achieve the best of each while overcoming their unique deficiencies. (p.302)

More specifically, in this study the results of classifying items on the basis of their MH alpha estimates were compared to the results of classifying these items using an IRT-based procedure involving the comparison of item difficulties (b-values) in the interest of identifying that alpha value that maximized the decision concordance between the two methrds.

Method

Description of Sample and Test Data

The data analyzed consisted of candidate responses to 299 scored items on an examination for professional licensure prepared by CTB/McGraw-Hill. A total of 68,458 candidates took this examination in 1988. The candidates' ethnicity was determined by their responses to a demographic question in which they were asked to classify themselves in terms of one of seven ethnic categories. A total of 47,573 candidates classified themselves in Ethnic Group 1; 6,486 candidates classified themselves in Ethnic Group 2; 5,466 in Ethnic Group 3; 2,004 in Ethnic Group 4; 1,014 in Ethnic Group 5; 486 in Ethnic Group 6; 307 in Ethnic Group 7. An additional 726 candidates classified themselves as "Other", and the remaining 4,396 candidates did not specify their ethnicity.

For the Mantel-Haenszel analyses, all candidates who classified themselves were used. For the IRT analyses to be described, a random sample of 500 candidates was drawn from Ethnic Group 1, the majority group, and random samples of 1000 candidates were drawn from Ethnic Groups 2, 3, and 4, the three largest minority groups. All candidates who classified themselves in Ethnic Groups 5, 6, and 7 or as Other were used in the analyses.

Procedure

Both the Mantel-Haenszel and IRT-based analysis to be described entail comparisons of the item performance of two groups, a reference group and a focal group. In this study, the candidates in Ethnic Group 1, the majority group, were used as the reference group, and the seven other candidate groups each were designated a focal group.

Mantel-Haenszel Analysis. To use the Mantel-Haenszel (MH) procedure, the reference group and focal groups are matched on ability. To match the groups, the total test score typically is used (Holland & Thayer, 1986) to sort examinees into score groups. Candidates in a given score group are then classified in terms of whether they answer each item correctly or incorrectly. In Figure 1, this type of classification of responses to an item is shown for score group j.

Insert Figure 1 about here

In Figure 1, T_j refers to the total number of candidates in score group j; A_j, B_j, C_j , and D_j to the number of candidates in each of the four cells; and T_{1j} , T_{0j}, T_{Rj}, and T_{Fj} to the marginals.

The MH estimate of alpha (α), which expresses the common odds ratio of success of the two groups across all score groups, can be defined as

$$\alpha = \frac{\Sigma A_j D_j / T_j}{\Sigma B_j C_j / T_j} . \qquad (1)$$

Alpha can vary between 0 and ∞ . As noted previously, when $\alpha = 1$ the odds for success are the same in the reference group and focal group. That is, the reference group and the focal group demonstrate the same performance on an item. This finding indicates that nc differential item functioning (DIF) or potential bias has been detected.

In this study, candidates' total raw scores on the examination were used to divide the candidates into score groups. A total of 13 score groups were constructed for the analyses comparing the majority group, Ethnic Group 1, with the four largest minority groups, Ethnic Groups 2 through 5. At least 50 candidates fell in each of these 13 score groups, which had the following raw score ranges: 135-143, 144-152, 153-156, 157-161, 162-168, 169-171, 172-175, 176-178, 179-182, 183-187, 188-194, 195-202, and 203-221. Nine score groups were constructed for the analyses comparing Ethnic Group 1 with the three smallest candidate groups, Ethnic Groups 6, 7, and Other. Fewer score groups were used for these analyses in order to ensure that there were enough candidates in each score group for adequate matching to occur. At least 22 candidates fell in each of these nine score groups, which had the following ranges: 157-165, 166-173, 174-181, 182-186, 187-193, 194-199, 200-203, 204-211, and 212-221. Candidates scoring below and above the listed raw score ranges were excluded from the analysis because their disparate scores and low counts would not permit adequate matching.

The IRT-Based Procedure

The IRT-based method used in the study was derived from Lord (1980, pp. 219-220), and it entailed four steps. First, the Rasch item difficulty (b-parameter) for each item was estimated using LOGIST 5 (Wingersky, Barton, & Lord, 1982) and the sample of 500 candidates randomly drawn from the majority reference group. Rasch ability estimates for these candidates generated by LOGIST were standardized, that is, scaled to have a mean of 0.0 and a standard deviation of 1.0. Then item difficulties were re-estimated seven times, using the responses of one of the seven focal groups for each estimation. The seven sets of item difficulties that were estimated for the seven focal groups subsequently were rescaled to place these sets on the same scale as that of the item difficulties estimated for the reference group. Finally, to make each of the seven reference group-focal group comparisons of item performance, a t-statistic for each item was calculated to assess the difference between the difficulty of the item for the reference group and for the focal group. The t-statistic is expressed as

$$\frac{\hat{b}_{Ri} - \hat{b}_{Fi}}{\surd(\text{var } \hat{b}_{Ri} + \text{var } \hat{b}_{Fi})},$$

where $\hat{b}_{Ri}$ and $\hat{b}_{Fi}$ are the estimated item difficulty of item i for the reference group and the focal group, respectively, and var $\hat{b}_i$ is expressed as

$$\frac{1}{\mathcal{E}\left[\left(\frac{\partial \ln L}{\partial \hat{b}_i}\right)^2\right]}.$$

Procedure for Identifying Cutscores

For each reference group-focal group comparison, the MH estimates of alpha obtained for all 299 items in the examination were plotted against the t-statistics obtained for these items in a bivariate plot. Subsequently the bivariate plot was partitioned by selecting a MH cutscore and a t-statistic cutscore, appearing on the x and y axis, respectively, and drawing perpendicular lines from the axes at these cutscores through the plot; the intersection of these lines created four quadrants. Quadrant 1 contained items with alphas and t-statistics that were greater than or equal to the selected alpha and t-statistic cutscores, respectively. These items will be referred to as "potentially biased" in this paper. Quadrant 3 contained items with both alphas and t-statistics less than the selected alpha and t-statistic cutscores, respectively. Thus Quadrants 1 and 3 contained items that were consistently classified on the basis of the two cutscores. In contrast, Quadrant 2 contained items with alphas greater than or equal to the selected alpha cutscore and t-statistics less than the selected t-statistic cutscore; Quadrant 4 contained items with alphas less than the selected alpha cutscore and t-statistics greater than or equal to the selected t-statistic cutscore. Thus, Quadrants 2 and 4 contained items that were inconsistently classified on the basis of the two cutscores.

As one measure of concordance between the results of the MH and IRT-based procedures, counts were made of the number of items falling in Quadrants 1 and 3 combined. By selecting different MH values and t-statistic values and counting the items in the two quadrants, each plot was searched systematically using a computer algorithm to find the combination MH value and t-statistic value that produced the largest number of items falling in the two quadrants combined. On the alpha scale, the cutscores tried were 1.40, 1.41, 1.42, through 2.90 in steps of 0.01. Each of these cutscores was tried in combination with t-statistic cutscores of 1.96, 1.97, 1.98 through 7.56[1] in steps of 0.01, and the sum of the items falling in Quadrant 1 plus Quadrant 3 for each pair of cutscores was calculated. The ratio of this sum to the total number of items analyzed was called the concordance proportion.

Constraints were imposed on the algorithm used to identify the pair of cutscores that produced the maximum count of items falling in Quadrant 1 plus Quadrant 3. These constraints were necessary because a concordance proportion of 1.00 could be obtained simply by making the alpha and t-statistic cutscores so extreme that all items analyzed fell in Quadrants 1 or 3. The first set of constraints specified that when a new pair of cutscores was selected at least three items should be found in Quadrant 1[2], at least two items should be found in Quadrant 2, and at least two items in Quadrant 4. This set of modest constraints, which required that about 1% of the items appear in Quadrant 1 and slightly less than 1% appear in Quadrant 2 and Quadrant 4, served to prevent the selection of extreme cutscores. The second constraint specified that when a new pair of cutscores produced a higher concordance proportion than a previous pair, the new pair could be considered a new maxima only when this new pair produced a proportional loss of items in Quadrant 1 that was no

[1] These t-statistics correspond to probabilities that range from $p = .006$ to $p < .000001$, one-tailed.

[2] A baseline comparison conducted by randomly assigning the members of the majority group to two comparison groups indicated that at a $p < .01$ significance level five items would be flagged as potentially biased by chance alone using the t-statistic employed in the IRT method. A similar baseline comparison using the MH chi-square test indicated that 0 items would be flagged on the basis of chance alone.

greater than the proportional loss in erroneously classified items in Quadranc 2 plus Quadrant 4. In effect, this second constraint insured that the increase in the number of items in Quadrant 3 had to be due to reductions in erroneous classifications (in Quadrant 2 plus Quadrant 4) that were at least as great as any reductions in the number of Quadrant 1 items. In the event that more than one pair of cutscores produced the maximum concordance proportion, the pair of cutscores consisting of the smallest alpha and t-statistic was selected as the maximizing cutscore.

Results

In Table 1 are shown the percentage of candidates in each of the 13 raw score groups constructed to conduct the MH analyses on the five largest ethnic groups. In Table 2 are shown the percentage of candidates in each of the nine raw score groups constructed to conduct the MH analyses on the three smallest ethnic groups. Also noted in these two tables are the percentage of candidates in each ethnic group who had extreme raw scores and were excluded from the MH analyses.

The data in Tables 1 and 2 show that candidates in Ethnic Groups 2, 5, and 6 were fairly evenly distributed over the score groups included in the analyses. The scores for candidates in Ethnic Groups 1, 3, 4, and 7 were more heavily concentrated in the higher raw score groups. The scores for candidates in the Other group were somewhat more heavily concentrated in the lower raw score groups.

Table 3 provides summary statistics describing the alpha and t-statistic values computed for each of the seven reference-focal group comparisons that were analyzed. With respect to the alpha values, the table shows that the reference-focal group comparisons having low mean alphas also had low variability in their alpha values. In contrast reference-focal group comparisons having relatively higher mean alphas also had higher variability in their alpha values. Specifically, low mean and median alpha values between .98 and 1.10 were obtained for the comparisons involving Ethnic Group 1 vs. Ethnic Groups 3, 4, 6, 7, and Other. Relatively low variability in the range of alpha values also were found in these comparisons, suggesting low levels of DIF as measured by the MH alpha were evident in these comparisons. Notably higher mean alphas and greater variability were found for the comparisons involving Ethnic Group 1 vs. Ethnic Groups 2 and 5. The median alpha values for these two comparisons were very close to 1.00 and lower than the means, indicating positive skews in the distribution of alpha values, suggesting the presence of some items with higher levels of DIF in these two comparisons.

With respect to the t-statistics obtained for the seven reference-focal group comparisons, a similar pattern of findings was observed. Specifically, mean t-statistics between -.03 and -.14 were obtained for the comparisons involving Ethnic Group 1 vs. Ethnic Groups 3, 4, 6, and 7, and the variability in the t-statistics obtained for these comparisons was also relatively low. Thus, low levels of DIF as measured by the t-statistic appear to be evident in these four comparisons, as was the case when their levels of alpha were analyzed. Notably more negative mean t-statistics and greater variability were observed for the comparisons involving Ethnic Group 1 vs. Ethnic Groups 2, 5, and Other. In all comparisons except Ethnic Group 1 vs. Ethnic Groups 3 and 6, the median t-statistics were somewhat higher than the means. This suggests somewhat negative skews in the distribution of these statistics obtained for each comparison, particularly for the comparisons of Ethnic Group

1 vs. Ethnic Groups 2 and 5. For these two comparisons, it appears again that there are some items with higher levels of DIF as measured by the t-statistics.

In Table 4 are the correlations between the alpha values and t-statistics calculated for each of the seven comparisons. These correlations were consistently high, ranging from .75 to .84, suggesting a consistently strong and positive relationship between the values produced by the MH and IRT methods.

Bivariate plots of the alpha and t-statistics calculated for the 299 items and seven reference-focal group comparisons are provided in Figures 2 through 8. It should be noted that because of constraints inherent in the plotter, the point placement must be regarded as approximate. In general these plots show a somewhat curvilinear relationship between the alpha and t-statistics calculated for each comparison, which is the expected relationship between two variates, one of which is in an antilog relationship to the other. It should be noted that there was evidence of a more pronounced curvilinear relation between the two statistics in Figures 2 and 5, which involved the comparisons of Ethnic Group 1 vs. Ethnic Groups 2 and 5.

Provided in Table 5 for each of the seven comparisons are the cutscores that maximized the proportion of concordant ratings resulting from use of the two methods. These cutscores have also been drawn on the plots in Figures 2 through 8. The maximizing alpha cutscores varied somewhat across the seven comparisons, ranging between 1.56 to 1.99. The maximizing t-statistic cutscores varied more substantially, ranging from 2.52 to 5.92. A comparison of the maximizing cutscores obtained for each reference-focal group comparison with the alpha and t-statistics reported in Table 3 shows a pattern: Lower cutscores were derived for those comparisons with lower and less variable leve , of DIF, and higher cutscores were derived for those comparisons that appeared to have some items with more substantial amounts of DIF.

Table 6 shows for each reference-focal group comparison the distribution of items in the four plot quadrants at the maximizing cutscores and measures of the degree of concordance between the item classifications resulting from use of these cutscores. With respect to the distributions of items across the four quadrants, note should be made of the numbers of items that appeared in the first quadrant across the seven comparisons. Table 6 shows that these numbers ranged from four items tc 52 items, with the highest numbers of items observed for the comparisons involving Ethnic Group 1 vs. Ethnic Groups 2 and 5. It was these comparisons that both the alpha and t-statistics independently suggested had items with higher levels of DIF.

Two measures of concordance for the seven comparisons of interest are reported in Table 6. The first measure, called the concordance proportion (p_o) in this paper, was first suggested by Hambleton and Novick (1974) as a measure of the consistency of mastery/non-mastery classifications. In the current study it was used to assess the proportion of items that were consistently classified by the two methods at the maximizing cutscores. Table 6 shows that the maximizing cutscores generally produced very high concordance rates, which ranged between .94 and .99.

The second measure, Cohen's (1960) kappa (k) has been recommended in the literature (Swaminathan, Hambleton, & Algina, 1974) as a useful index of decision consistency. It indicates the degree to which the proportion of decisions found to be consistent over methods exceeds the proportion to be expected by chance, that is, when the two decision methods are statistically independent. As has been noted by Subkoviak (1980) and Traub and Rowley (1980), the relationship between the p_o and k is complex because the two statistics are affected differently by (1) the shapes of the two distributions of scores being analyzed, (2) the location of the cutscores in these two distributions, and (3) the correlations between two distributions of scores.

The results for k reported in Table 6 were somewhat lower than those reported for p_o, although with the exception of the Ethnic Group 1 vs. Ethnic Group 4 comparison, the k values were high. The lower value of k for this comparison appears to be due to the unusual degree of misclassification that is apparent when the numbers of items in plot quadrants 1 and 2 are comoared. As the table shows, of the 10 items that exceeded the maximizing alpha cutscore, only four of these items also exceeded the maximizing t-statistic cutscore and hence fell in plot quadrant 1; the remaining six items fell below the maximizing t-statistic cutscore and fell in quadrant 2.

Table 7 provides a summary of selected information presented in Tables 3 and 6. In this table, the reference-focal group comparisons were first ranked from low to high in order of the value of the MH alpha at the third quartile in the distributions of alpha values calculated for the seven comparisons. The comparisons were also ranked from low to high in order of the value of the t-statistic at the third quartile in the distributions of t-statistics calculated for the seven comparisons. As is evident the rank orders of the comparisons based on the two statistics were nearly identical, indicating that the statistics calculated using the two methods similarly distinguished between the seven comparisons in terms of the degree of potential bias in the items analyzed. These findings corroborate the high correlations reported in Table 4. The similarity between the ranks of the third quartile values of the two statistics and the ranks of the number of items in the first quadrant of the plot further indicated that the two methods produced cutscores that ordered the comparisons similarly in terms of their potential bias. That is, comparisons independently identified by the two methods as having low levels of potential bias (i.e., low alphas and t-statistics at the third quartile) also were found to have few items in the first quadrant. In contrast, comparisons independently identified by the two methods as having relatively higher levels of potential bias (i.e., relatively high alphas and t-statistics at the third quartile) were found to have substantially more items in the first quadrant.

Finally, Table 8 shows what results would obtain were significance levels for each of the two methods used to detect DIF were used in lieu of a cutscore derived using the multiple method approach described in this paper. Using the MH chi-square test, between 11 and 132 items would be flagged for DIF against a minority group at p <.01. If a significance level of p <.01 for the t-statistics were used, between 5 and 94 items would be flagged for this DIF. In contrast, as shown in Table 6, between 5 and 52 items were flagged for DIF using the multiple method approach. It should also be noted that the numbers of items shown to be flagged for DIF in Table 8 under each method were strongly related to the sizes of the minority groups that were analyzed in the seven comparisons, whereas the numbers of items flagged for DIF using the maximizing cutscores were not related to sample size.

Discussion

The findings suggest that the use of a multiple method approach to define an alpha criterion for users of the MH method has advantages over the more traditional methods of using a significance level to establish this criterion. More specifically, the use of the multiple methods enabled the identification of items as potentially biased through a concordance of results from two independent methods of assessing DIF. Presumably this approach permits greater generalizability of the findings, as well as greater confidence that the classification of items is accurate.

The IRT method used in conjunction with the MH method involved assessing DIF by computing a t-statistic for each item that indicated the degree to which the difficulty of the item differed for the reference and focal group. As should be expected of an alternative method of assessing uniform DIF, the t-statistics resulting from use of this method correlated highly with the MH alpha values and ordered the seven reference-group comparisons similarly in terms of the magnitude of DIF detected. When the IRT method was used in conjunction with the MH analyses, cutscores were identified that produced a high level of decision concordance.

It is evident that the maximizing alpha cutscores may vary over the reference-focal group comparisons. In the current study, these cutscores ranged from 1.56 for the comparison of Ethnic Group 1 vs. Ethnic Group 7 to 1.98 and 1.99 for the two comparisons of Ethnic Group 1 vs. Ethnic Groups 2 and 5, respectively. It is interesting to note that the former comparison produced the lowest mean alpha, smallest standard deviation of alpha, and the least skew in its distribution of alpha values; the latter two comparisons produced the highest means and standard deviations and greatest positive skews. The t-statistics for these two comparisons showed similar discrepant patterns. These findings suggest that varying degrees of DIF affect the minority groups involved in these comparisons, and that the nature of the DIF detected has a differential impact on the relationship between the alpha and t-statistics and, hence, on the maximizing cutscores.

One possible explanation for the different maximizing cutscores is that different factors ar producing the differential in potential bias evident across comparisons. For example, cultural differences may be active in some comparisons, whereas in others curriculum effects are also present. It is noteworthy that Ethnic Groups 2 and 5 are known to be comprised of largely foreign-educated candidates.

If one alpha cutscore must be used where multiple comparisons are to be made, it seems most reasonable to choose the lowest of the maximizing cutscores that are identified for the reference-focal groups being analyzed. Setting the cutscore at the lowest of these values ensures that any misclassification errors accrued by using a "non-optimal" cutscore for some of the comparisons would not be disadvantageous to the minority groups in these comparisons. Using the "non-optimal" cutscore would result in more erroneous classifications of items as having potential bias, but no items with DIF would fail to be classified because of this cutscore.

The methodology employed in this study has been replicated on another examination for professional licensure prepared by CTB/McGraw-Hill, where the candidate population includes some of the same ethnic groups involved in the current study. For the reference-focal group comparisons that were the same, the findings from this replication were comparable to those in the current study. That is, high levels of decision concordance were found, and those comparisons with higher levels of DIF as measured by the alpha and t-statistics also were found to have more items classified in the first plot quadrant. The numbers of items in this quadrant also did not appear to be a function of sample size. Furthermore, in those comparisons involving predominantly foreign-educated candidates in the focal group, more extreme alphas and t-statistics as well as more items in the first plot quadrant were found.

The merits of the proposed methodology rests on its generalizability and accuracy. It is recommended that the methodology be replicated in the future using other examinations and other reference-focal group comparisons. In addition, simulation studies should be done to assess the degree to which the proposed methodology effectively distinguishes between items that do and do not possess "true" DIF.

References

Berk, R. A. (Ed.) (1982). Handbock of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Denzin, N. K. (1978). The research act: A theoretical introduction to sociological methods. New York: McGraw-Hill.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.) Educational Measurement (3rd edition). New York: American Council on Education/Macmillan Publishing Co., pp. 201-219.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Hambleton, R. K., & Rogers, J. R. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.

Hills, J. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.

Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd edition). New York: American Council on Education/Macmillan Publishing Co., pp. 13-103.

Phillips, A., & Holland, P. W. (1937). Estimators of the variance of the Mantel-Haenszel log-odds ratio estimate. *Biometrics*, *43*, 425-431.

Subkoviak, M. J. (1980). Estimating the reliability of mastery-non mastery classifications. In R. A. Berk (Ed.) *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins University Press, pp. 267-291.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. *Journal of Educational Measurement*, *11*, 263-267.

Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement*, *4*, 517-545.

Wingersky, M. S., Barton, M. A., & Lord, F. M. *LOGIST 5.0 version 1.0 users' guide*. Princeton, NJ: Educational Testing Service. (Version 2.5 updated 1984).

Table 1

Percentage of Candidates in Each of 13 Raw Score Groups
and Excluded Raw Score Ranges
by Ethnic Group

	Raw Score Group														
Ethnic Group	<134*	135-143	144-152	153-156	157-161	162-168	169-171	172-175	176-178	179-182	183-187	188-194	195 202	203-221	>222*
1	0.2	0.2	0.5	0.4	0.8	2.1	1.2	2.2	2.0	3.4	5.3	9.2	12.9	32.9	26.8
2	18.7	6.9	8.8	4.5	6.1	10.1	4.1	5.6	4.5	5.4	6.1	6.9	5.2	5.5	1.4
3	2.8	2.2	4.0	2.8	3.9	8.9	4.4	6.4	5.2	7.4	8.6	11.7	11.9	15.3	4.4
4	6.4	2.8	4.1	2.5	4.0	6.4	3.6	4.3	4.4	5.4	7.5	11.2	10.6	19.0	7.7
5	16.5	7.4	10.0	5.5	5.8	10.2	4.9	6.0	6.0	5.3	5.6	5.8	4.9	5.0	0.9

* Cases in this score range were excluded from the analyses.

Table 2

Percentage of Candidates in Each of 9 Raw Score Groups
and Excluded Raw Score Ranges
by Ethnic Group

	Raw Score Groups										
Ethnic Group	<157*	157-165	166-173	174-181	182-186	187-193	194-199	200-203	204-211	212-221	>221*
1	1.2	1.8	3.3	5.7	5.2	8.8	9.5	6.7	14.3	16.7	26.8
6	18.3	9.5	10.1	10.9	7.8	8.6	4.7	5.1	8.0	5.1	11.7
7	4.6	7.2	7.2	7.8	8.5	8.5	10.7	7.8	9.8	11.7	16.3
Other	34.8	8.1	10.0	10.3	5.0	6.1	5.0	3.3	4.4	5.5	7.4

* Cases in this score range were excluded from the analyses.

Table 3

Summary Statistics Describing Alpha and T-Statistic Values by Reference-Focal Group Comparison

Reference-Focal Group Comparison	Alpha Values					T-Statistic Values				
	Mean	S.D.	Median	Q3	Q1	Mean	S.D.	Median	Q3	Q1
Ethnic Group 1 vs. Ethnic Group 2	1.20	.85	.98	1.47	.67	-.23	4.93	.09	3.04	-3.49
Ethnic Group 1 vs. Ethnic Group 3	1.06	.33	1.04	1.27	.82	-.14	2.80	-.21	1.90	-2.03
Ethnic Group 1 vs. Ethnic Group 4	1.04	.26	1.01	1.17	.86	-.14	2.31	.10	1.50	-1.79
Ethnic Group 1 vs. Ethnic Group 5	1.29	1.08	1.00	1.65	.61	-.31	5.60	.04	3.39	-4.20
Ethnic Group 1 vs. Ethnic Group 6	1.08	.45	.98	1.27	.78	-.12	3.00	-.14	1.94	-2.27
Ethnic Group 1 vs. Ethnic Group 7	1.03	.23	1.02	1.14	.88	-.03	1.31	.08	.80	-.90
Ethnic Group 1 vs. Other	1.10	.49	1.02	1.33	.77	-.19	3.55	-.07	2.44	-2.58

Table 4

Correlations between Mantel-Haenszel Alphas and
T-Statistic Values by Reference-Focal Group Comparison

Reference-Focal Group Comparison	r_{xy}
Ethnic Group 1 vs. Ethnic Group 2	.78
Ethnic Group 1 vs. Ethnic Group 3	.84
Ethnic Group 1 vs. Ethnic Group 4	.79
Ethnic Group 1 vs. Ethnic Group 5	.78
Ethnic Group 1 vs. Ethnic Group 6	.80
Ethnic Group 1 vs. Ethnic Group 7	.75
Ethnic Group 1 vs. Other	.80

Table 5

Cutscores that Maximize the Concordant Classifications by Reference-Focal Group Comparison

Reference-Focal Group Comparison	Concordance Maximizing Alpha Cutscore	Concordance Maximizing T-statistic Cutscore
Ethnic Group 1 vs. Ethnic Group 2	1.98	5.92
Ethnic Group 1 vs. Ethnic Group 3	1.73	4.63
Ethnic Group 1 vs. Ethnic Group 4	1.59	3.88
Ethnic Group 1 vs. Ethnic Group 5	1.99	3.75
Ethnic Group 1 vs. Ethnic Group 6	1.76	4.02
Ethnic Group 1 vs. Ethnic Group 7	1.56	2.52
Ethnic Group 1 vs. Other	1.93	4.82

Table 6

Counts of Items in Four Plot Quadrants, Total Concordance and Concordance Statistics at Maximizing Cutscores by Reference-Focal Group Comparison

Reference-Focal Group Comparison	Plot Quadrant				Total Condordance	Concordance Proportion (p_o)	Kappa (k)
	1	2	3	4	(1 + 3)	(1 +3)/299	
Ethnic Group 1 vs. Ethnic Group 2	30	10	253	6	283	.95	.76
Ethnic Group 1 vs. Ethnic Group 3	7	2	288	2	295	.99	.77
Ethnic Group 1 vs. Ethnic Group 4	4	6	287	2	291	.97	.49
Ethnic Group 1 vs. Ethnic Group 5	52	3	228	16	280	.94	.81
Ethnic Group 1 vs. Ethnic Group 6	16	4	274	5	290	.97	.76
Ethnic Group 1 vs. Ethnic Group 7	5	2	290	2	295	.99	.71
Ethnic Group 1 vs. Other	17	3	272	7	289	.97	.76
Total Across Groups	131	30	1892	40	2023	.97	.77

Table 7

Reference-Focal Group Comparisons Ranked[1] by
Mantel-Haenszel and T-Statistic Values at the Third Quartile (Q3) and
by the Number of Items in the First Quadrant of Plot

Reference-Focal Group Comparison	Mantel-Haenszel Alpha at Q3		T-Statistic at Q3		Items in First Quadrant of Plot	
	Value	Rank	Value	Rank	Number	Rank
Ethnic Group 1 vs. Ethnic Group 7	1.14	1	.80	1	5	2
Ethnic Group 1 vs. Ethnic Group 4	1.17	2	1.50	2	4	1
Ethnic Group 1 vs. Ethnic Group 3	1.27	3.5	1.90	3	7	3
Ethnic Group 1 vs. Ethnic Group 6	1.27	3.5	1.94	4	16	4
Ethnic Group 1 vs. Other	1.33	5	2.44	5	17	5
Ethnic Group 1 vs. Ethnic Group 2	1.47	6	3.04	6	30	6
Ethnic Group 1 vs. Ethnic Group 5	1.65	7	3.39	7	52	7

[1] Rank orders range from a low of 1 to a high of 7.

Table 8

Number of Significant (p< .01) Mantel-Haenszel Chi-Square Values and Significant T-Statistics Compared to Number of Items in Plot Quadrant 1 by Reference-Focal Group Comparison

	Significant Mantel-Haenszel Chi-Squares			Significant T-Statistics			Number in Quad. 1
Reference-Focal Group Comparison	N	Total Number	Number Showing DIF Against Minority Group	N	Total Number	Number Showing DIF Against Minority Group	
Ethnic Group 1 vs. Ethnic Group 2	47,573 6,486	261	132	500 1,000	175	83	30
Ethnic Group 1 vs. Ethnic Group 3	47,573 5,466	221	118	500 1,000	106	54	7
Ethnic Group 1 vs. Ethnic Group 4	47,573 2,004	157	80	500 1,000	85	38	4
Ethnic Group 1 vs. Ethnic Group 5	47,573 1,014	223	113	500 1,014	189	94	52
Ethnic Group 1 vs. Ethnic Group 6	47,573 486	114	63	500 486	126	56	16
Ethnic Group 1 vs. Ethnic Group 7	47,573 307	22	11	500 307	16	5	5
Ethnic Group 1 vs. Other	47,573 726	140	75	500 726	142	67	17

Figure 1

Table of Frequencies Used in the Mantel-Haenszel Analysis of the Performance of a Score Group on an Item

	Item Score		
	0	1	
Reference Group	A_j	B_j	T_{Rj}
Focal Group	C_j	D_j	T_{Fj}
	T_{1j}	T_{0j}	T_j

Figure 2
Plot of t Statistics by M.H. Alpha
Ethnic Group 2

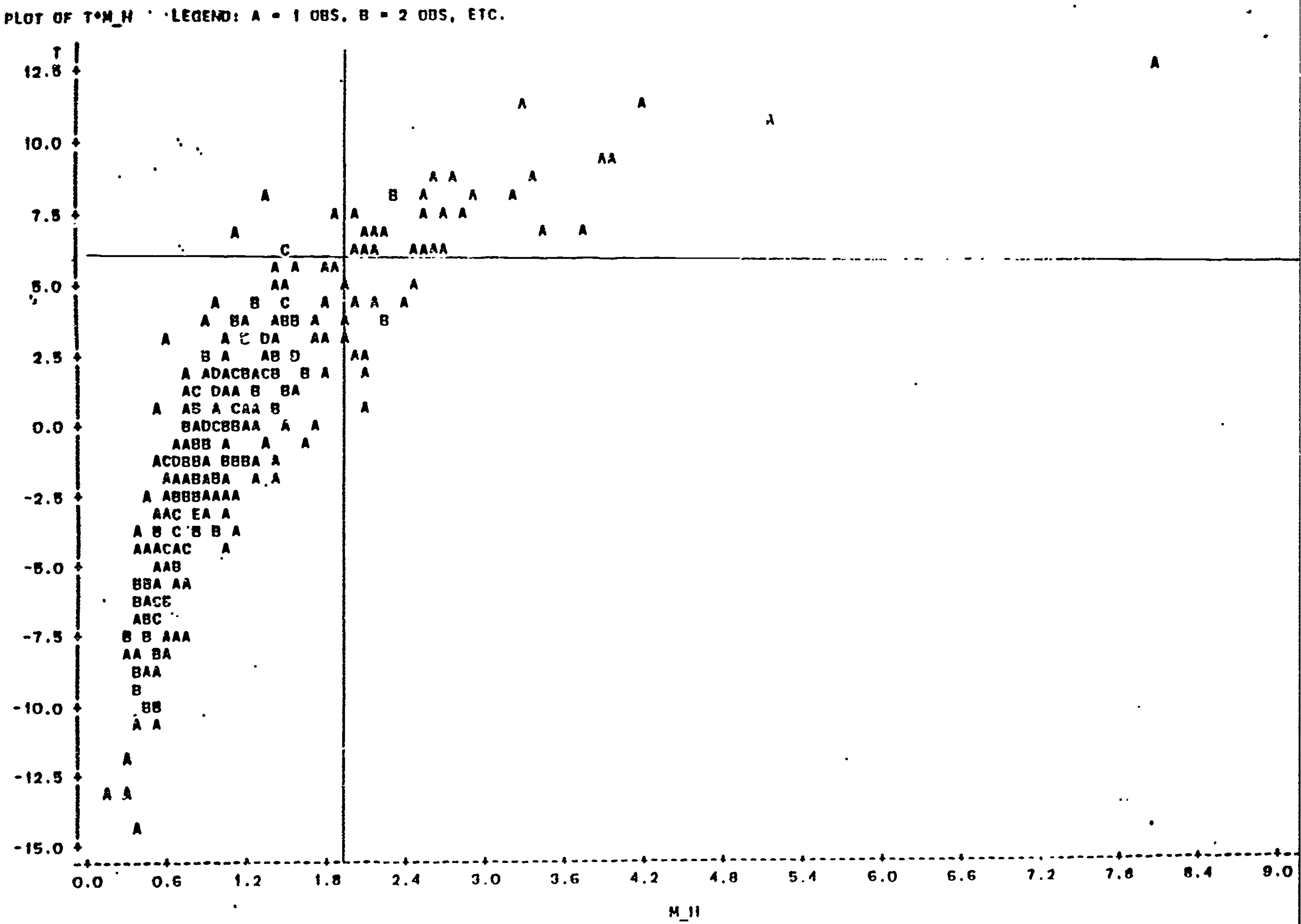


Figure 3
Plot of t Statistic by M.H. Alpha
Ethnic Group 3

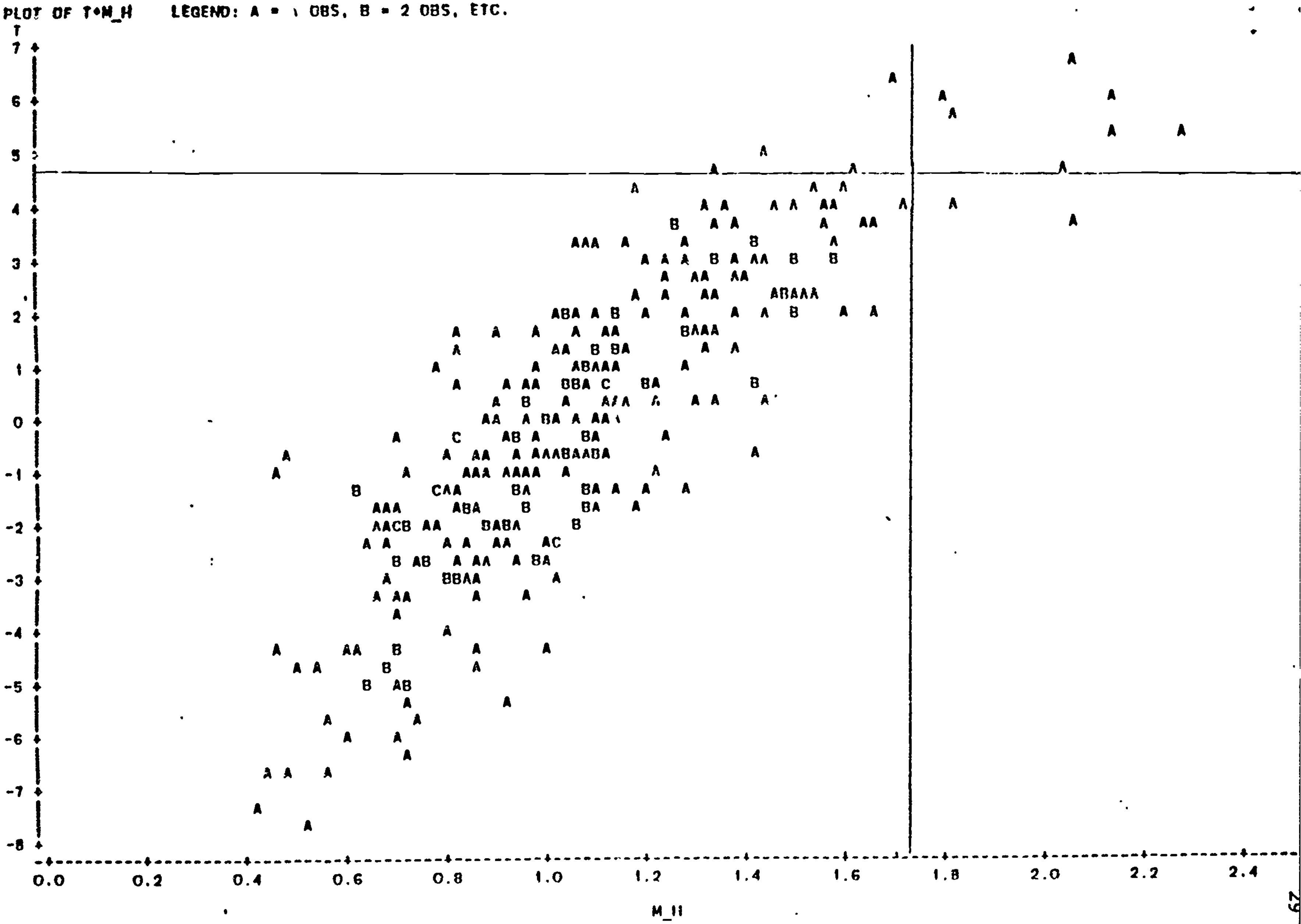


Figure 4
Plot of t Statistic by M.H. Alpha
Ethnic Group 4

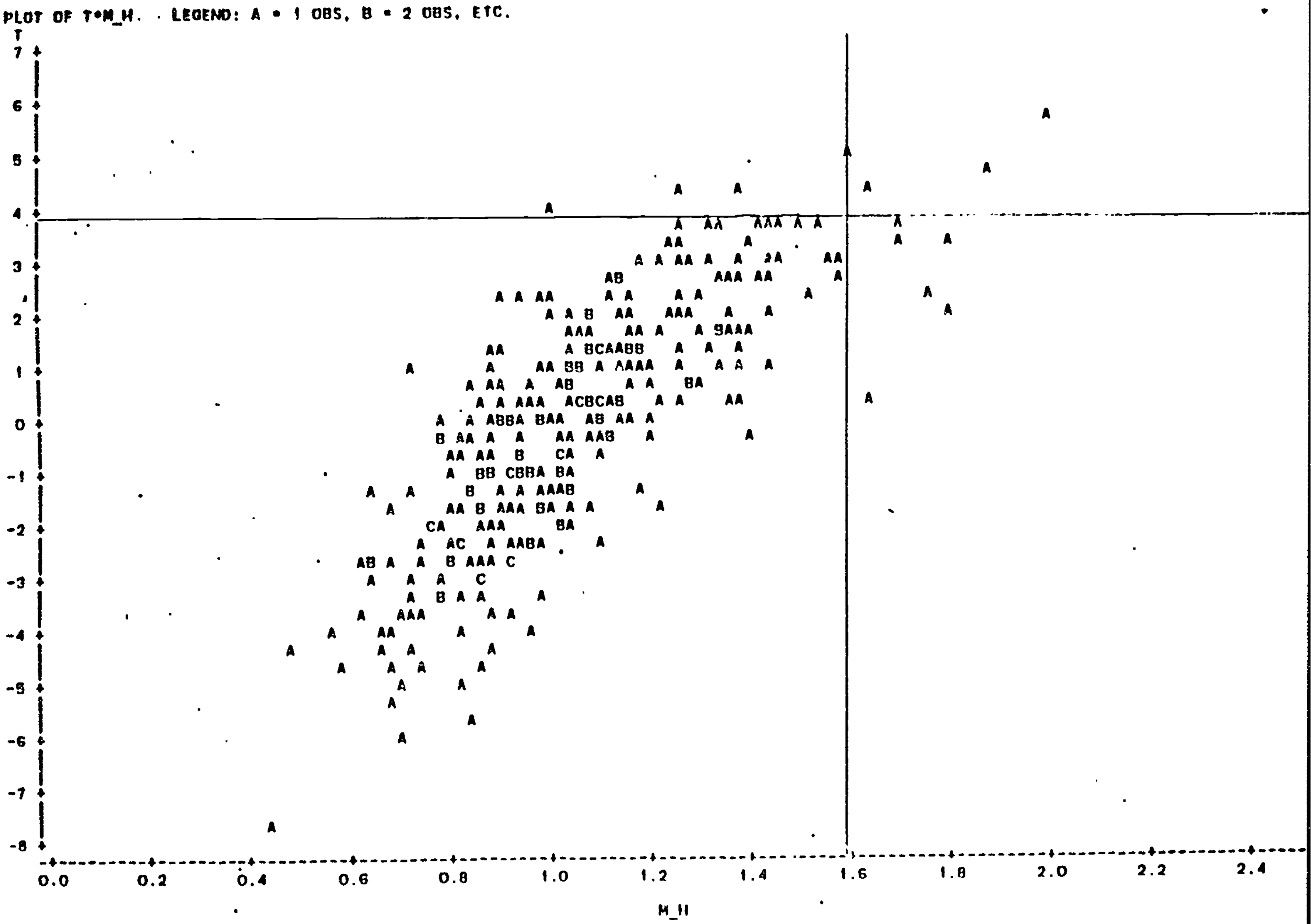


Figure 5
Plot of t Statistics by M.H. Alpha
Ethnic Group 5

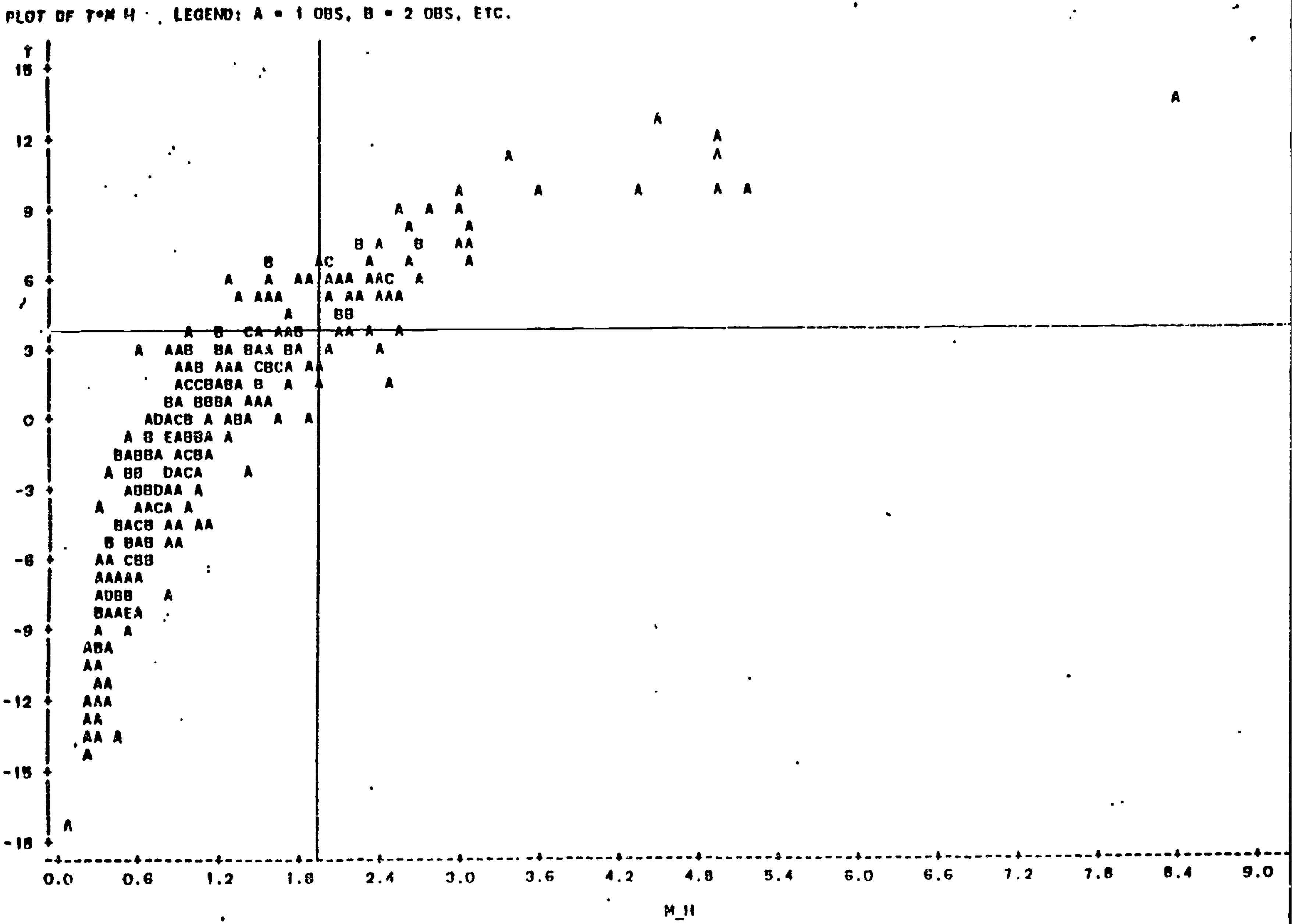


Figure 6
Plot of t Statistic by M.H. Alpha
Ethnic Group 6

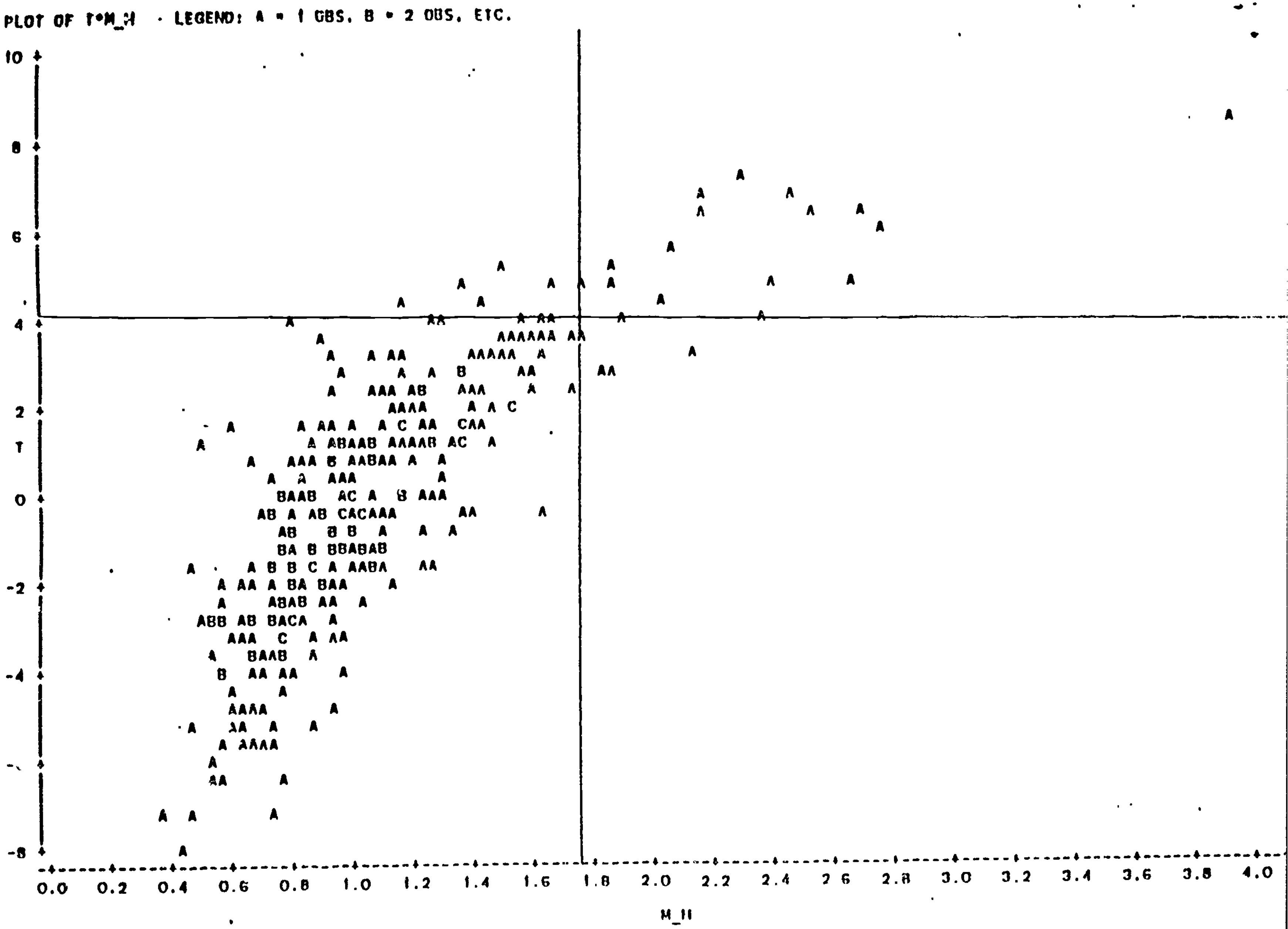


Figure 7
Plot of t Statistics by M.H. Alpha
Ethnic Group 7

PLOT OF T*M_H   LEGEND: A = 1 OBS, B = 2 OBS, ETC.

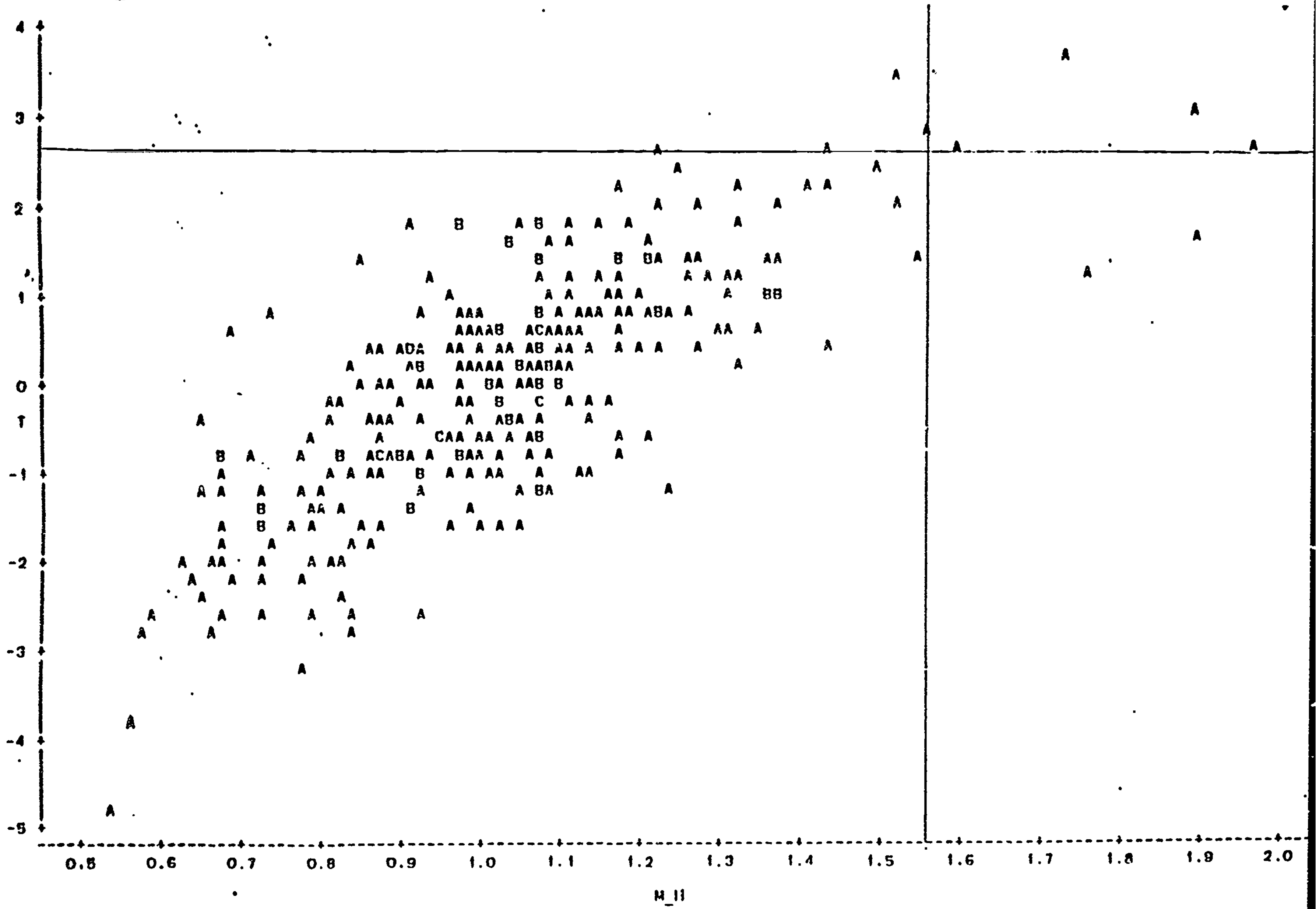
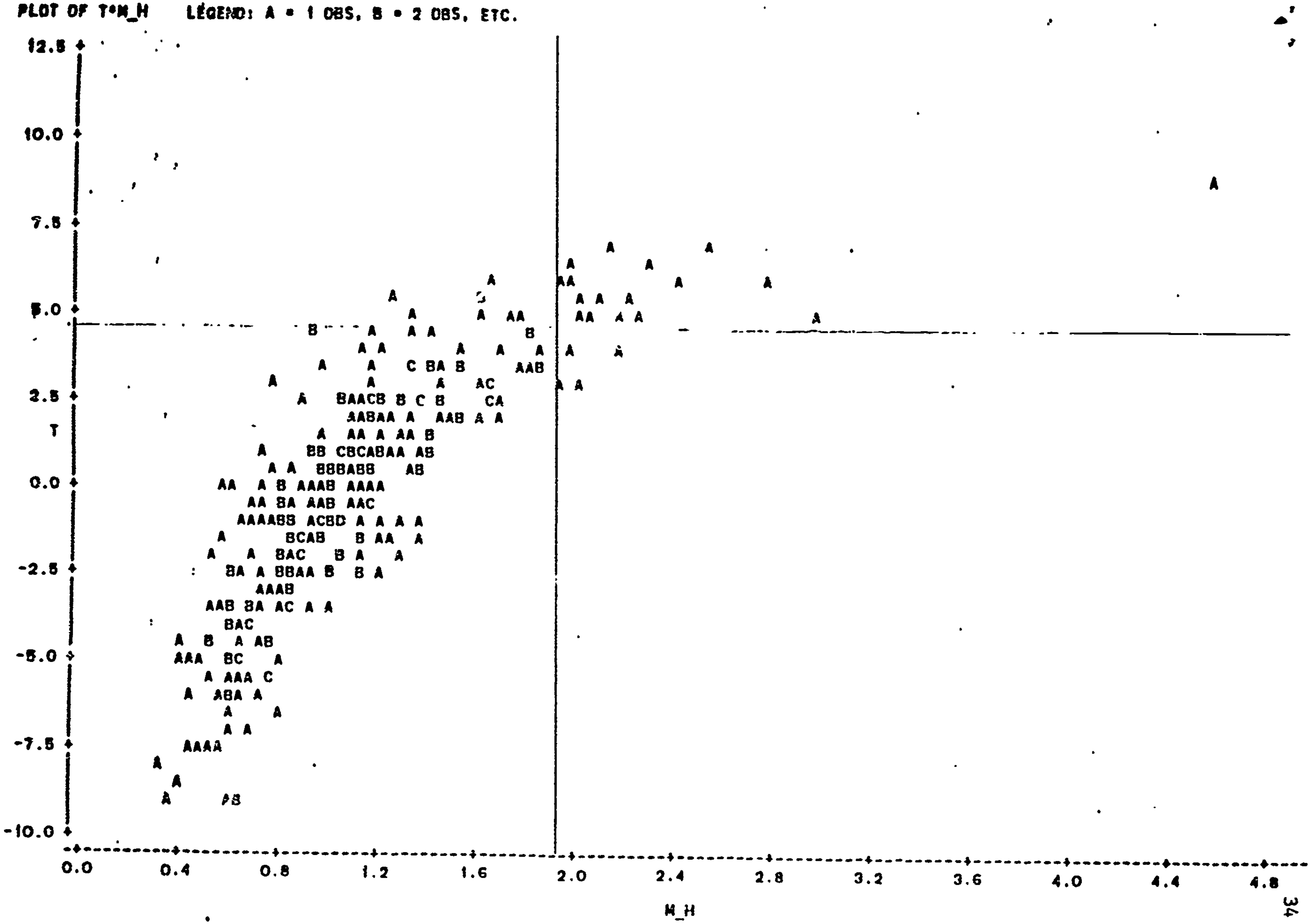


Figure 8
Plot of t Statistics by M.H. Alpha
Others


PLOT OF T*M_H LEGEND: A = 1 OBS, B = 2 OBS, ETC.
T
M_H