DOCUMENT RESUME

ED 419 833 TM 028 408

AUTHOR	Huang, Chi-Yu
TITLE	Factors Influencing the Reliability of DIF Detection Methods.
PUB DATE	1998-04-00
NOTE	25p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Identification; *Item Bias; *Regression (Statistics); *Reliability; *Sampling; Tables (Data)
IDENTIFIERS	Iowa Tests of Basic Skills; Item Bias Detection; Mantel Haenszel Procedure; Standardization

ABSTRACT

This study examined the reliability of three methods for detecting differential item functioning (DIF) (i.e., the Mantel-Haenszel method, the standardization method, and the logistic regression method) applied to achievement test data. In addition, the study examined the influences of different sources of error variance, including examinee, occasion, and curriculum sampling on the magnitude of the reliability of the different DIF detection methods. Three datasets were assembled from the 1992 spring and fall standardization administration of the Iowa Tests of Basic Skills, and these were manipulated to control for error variance sources. Results indicated that the Mantel-Haenszel and standardization methods were more reliable in detecting DIF than the logistic regression method. The data also indicated that controlling the error variance of curriculum sampling slightly increased the reliability of DIF detection while controlling for error variance due to examinee sampling gives confusing results. (Contains 13 tables, 8 figures, and 12 references.) (Author/SLD)

Factors Influencing the Reliability of DIF Detection Methods

Chi-Yu Huang
ACT, Inc.

Stephen B. Dunbar
The University of Iowa

Paper presented at the 1998 annual meeting of
the American Educational Research Association, San Diego, CA.

Factors Influencing the Reliability of DIF Detection Methods

Abstract

This study examined the reliability of three DIF detection methods (i.e., the Mantel-Haenszel method, the standardization method, and the logistic regression method) applied to achievement test data. In addition, the study examined the influences of different sources of error variance, including examinee, occasion, and curriculum sampling, on the magnitude of the reliability of the different DIF detection methods. Three datasets were assembled from the 1992 Spring and Fall standardization administration of the *Iowa Tests of Basic Skills* and were manipulated to control for error variance sources. Results indicated that the Mantel-Haenszel and standardization methods were more reliable in detecting DIF than the logistic regression method. The data also indicated that controlling the error variance of curriculum sampling slightly increased the reliability of DIF detection while controlling for error variance due to examinee sampling gave confusing results.

Introduction

According to Dorans and Holland (1993), differential item functioning (DIF) is defined as a psychometric difference in item performance between groups that are matched on the abilities or attributes measured by the test or items. Many methods have been developed to detect DIF. The Mantel-Haenszel (MH), standardization (STD), and logistic regression (LR) methods match examinees from various groups on observed test scores. These methods are different from IRT methods which detect differential functioning of items via matching examinees of different groups on estimated ability. The Mantel-Haenszel, standardization, and logistic regression methods have gained the attention of researchers and practitioners because of their straightforward definitions of DIF and easy implementation.

DIF analysis for test items is important in test development because it helps to examine and eliminate items that may be potentially unfair to subpopulations due to cultural or gender differences. If an item exhibits DIF during pretesting, the judgment of experts can be used to decide whether this item should be revised or deleted from consideration for the final form of the test. However, while reviewing these questionable items, experts often have difficulty in finding reasons to support the statistical results (Shepherd, Camilli & Williams, 1984; Skaggs & Lissitz, 1992); and results from expert and statistical procedures for detecting differentially functioning items have shown little agreement (Engelhard, Hansche & Rutledge, 1990; Hambleton & Jones, 1994; Plake, 1980; Qualls & Hoover, 1981). One possible reason for inconsistent results is that the DIF index derived from each analysis may not be as stable as it appears, despite a strong representation of samples to the target population (Hoover & Kolen, 1984). Therefore, examining the accuracy and the stability of DIF analyses is an important issue in the development and evaluation of DIF detection methods.

The present study focused on whether the MH, STD, and LR methods were sufficiently reliable to use in the detection of potentially biased items in a real testing situation. To improve the reliability of DIF detection methods, this study also controlled three sources of error variance which likely affect the magnitude of reliability in DIF detection. These three sources of error variance included examinee sampling, occasion sampling, and curriculum sampling. Examinee sampling was defined as error variance arising from differences in test responses on the same test items from different students who participated in the same test administration. Occasion sampling was defined as differences in test responses on the same test items from the same students when they were tested at different points in time. Curriculum sampling was defined as the variation due to the interaction between school curriculum and conditional group differences in item performance. The source of error variance from curriculum sampling was considered because previous studies of DIF methods have neglected to look at the possible contributing effect of differences in school curricula. Teachers use different methods and materials in their instruction, and differences in student performance may result from curriculum variance and not from differential item functioning. In the present study, three different datasets from the data pool from a national standardization administration of the *Iowa Tests of Basic Skills (ITBS)* in the Spring and Fall of 1992 were manipulated to control one or two sources of error variance at a time. Results from these different datasets were compared to provide answers to the following two questions: (1) which of the three DIF detection methods was more stable, and (2) which sources of error variance had a stronger effect on the magnitude of the reliability of DIF detection.

Methodology

Datasets

The data used in this study consisted of test results from the *Iowa Tests of Basic Skills* (*ITBS*) from fifth grade Caucasian and African-American students who were tested in the Spring of 1992 and from sixth grade students who were tested in the Fall of 1992. Three datasets were assembled to examine the impact of different sources of error variance on the reliability of DIF detection methods. Two subsets were included in each dataset and DIF analyses were performed separately for each subset. Table 1 delineates the three datasets (A through C) and the sources of error variance that were controlled in each dataset.

Table 1: Decomposition of the three datasets.

	Controlling Source of Error Variance in		
Dataset	Examinees	Occasions	Curricula
A		X	
B	X		
C		X	X

Dataset A included fifth grade students who took the same test form (Form K) in the Spring of 1992. This dataset was randomly divided into two subsets of students (A1 and A2) to examine the reliability of DIF detection methods when error variance from occasion sampling (time of test administration) was controlled.

Dataset B included students who took the same test form (Form K) of the *ITBS* in consecutive test levels during the Spring and Fall of 1992. The dataset included two subsets (B1 and B2). The first subset consisted of fifth graders who took Level 11 of the *ITBS* in the Spring of 1992. The second subset consisted of the same students from B1 who later took Level 12 in the sixth grade in the Fall of 1992. This dataset was used to examine the reliability of different DIF detection methods when the error variance from examinee sampling was controlled.

In Dataset C, Caucasian and African-American students who took Form K in the Spring were matched within school building to control the impact of error variance due to differences in curriculum across schools. The same number of Caucasian and African-American students was selected within each school building if both Caucasian and African-American students existed in the same building. Matched Students in each school were randomly divided into two subsets (C1 and C2) to evaluate the reliability of DIF detection methods when the effect of error variance from occasion and curriculum sampling were controlled. The sample sizes of Caucasian and African-American students for each subset are listed in Table 2.

Table 2: Sample sizes in three datasets.

	Dataset A		Dataset B		Dataset C	
	A1	A2	B1	B2	C1	C2
Caucasian	5,211	5,374	1,313	1,313	748	809
African-American	533	534	217	217	748	809

Procedure

DIF analyses were performed for common items in the Reading Comprehension, Spelling, Usage and Expression, and Math Computation tests of the *ITBS*. Caucasian students were the reference group and African-American students were the focal group for each analysis. Because Dataset B involved students who took tests at two consecutive levels (Levels 11 and 12), only common items in these two consecutive levels could be considered when the reliability of DIF detection methods was examined. To ensure that results from Dataset B were comparable with the other two datasets, DIF analyses in Datasets A and C also included only common items. There were 25 common items in the Reading Comprehension test, 21 items in the Spelling test, 20 items in the Usage and Expression test, and 20 items in the Math Computation test. The analyses using the MH, STD, and LR methods were performed separately for 24 subsets (2 subsets x 3 datasets x 4 tests).

In the MH analysis, index values of MH D-DIF and χ^2_{MH} were calculated for each item, and the DIF category for each item was determined based on classification rules developed by the Educational Testing Service (Dorans & Holland, 1993). Items were classified as exhibiting negligible DIF if the MH D-DIF value was not statistically different from zero, or if the magnitude of the MH D-DIF values was less than one delta unit in absolute value. Items were classified as exhibiting large DIF if the MH D-DIF exceeded an absolute value of 1.5 and was significantly larger than 1.0 in absolute value. All other items which did not fit the above criteria were classified as exhibiting intermediate DIF.

In the STD analysis, items were identified as exhibiting DIF based on the index of standardized P-difference (D_{STD}). According to the rules used by Dorans & Holland (1993), items were classified as exhibiting negligible DIF if the absolute values of D_{STD} were less than .05. If the absolute values of D_{STD} were between .05 and .10, items were classified as having intermediate DIF. Items with absolute values of D_{STD} greater than .10 were considered as exhibiting large DIF.

In the LR analysis, the chi-square value (χ^2_L) for the incremental effect of group membership and the interaction of ability and group membership served as the DIF index. Two types of DIF (i.e., non-uniform and uniform DIF) were considered when χ^2_L values significantly exceeded $\chi^2_{.05:2}$. If a chi-square value (χ^2_{NU}) for the interaction effect between ability and group membership significantly exceeded $\chi^2_{.05:1}$, items were classified as displaying non-uniform DIF. However, if an item was not classified as non-uniform but the chi-square value (χ^2_U) for the effect of group membership significantly exceeded $\chi^2_{.05:1}$, the item was classified as having uniform DIF. Items were classified as having no DIF if they did not satisfy the above criteria (Camilli & Shepard, 1994; Swaminathan & Rogers, 1990).

The reliability of DIF detection was assessed through the correlation analyses of DIF indexes and the degree of item classification inconsistencies between two subsets. Spearman's rank-order correlation was used in this study. Because most of the items in a test do not exhibit DIF, looking at the percent agreement of item classification does not provide a clear picture of how items in relation to DIF change between subsets. Therefore, instead of using the percent agreement as the indicator of reliability, the percent of the item classification inconsistencies was considered as another indicator of the reliability of DIF detection methods. Two levels of item classification inconsistency were defined in this study: serious inconsistency and minor inconsistency. For the MH and STD methods, serious item classification inconsistency was defined as item labels for the same item dramatically changing between "negligible DIF" and "large DIF" across subsets, and minor item classification inconsistency was defined as item classifications for the same item changing between "negligible DIF" and "intermediate DIF" across subsets, or when both of them were exhibiting DIF but having different degrees of DIF across subsets. For the LR method, serious item classification inconsistency was defined as item classifications for the same item dramatically changing from no DIF to exhibiting DIF (either uniform or non-uniform DIF) across subsets, and minor item classification inconsistency was defined when the same items were exhibiting DIF but changing between different types of DIF (i.e., uniform or non-uniform DIF) across subsets. Total item classification inconsistencies were calculated by adding the numbers of items identified as having either serious or minor inconsistencies.

Results and Discussion

The discussion of the results of this study is divided into two parts. Part one examines the reliability of the three DIF detection methods, with comparisons made within each dataset. There were twelve within dataset comparisons (3 datasets x 4 subtests). In the second part, comparisons of DIF analyses obtained from the three datasets based on the same detection method are presented to examine the influence of error variance of examinee, occasion and curricula sampling on reliability. For example, the reliability results of the MH method from Datasets A, B, and C for the Reading Comprehension test were compared to examine the influence of the three sources of error variance.

Comparison of the Reliability of Three DIF Detection Methods within Datasets

Correlation Analyses of DIF Indexes

Spearman correlation coefficients were calculated for each DIF index (MH D-DIF, χ^2_{MH}, D_{STD}, and χ^2_L) with themselves between the two subsets in each dataset (i.e., A1 with A2, B1 with B2, and C1 with C2) to examine the reliability of the three DIF detection methods. Table 3 through Table 6 list correlation results for each DIF index in the three datasets when common items in the Reading Comprehension, Spelling, Usage and Expression, and Math Computation tests were examined. Correlations for four DIF indexes were compared in twelve datasets (3 datasets x 4 subtests).

Insert Tables 3, 4, 5 and 6 here

Results from Table 3 through Table 6 show that, when correlation coefficients for the four DIF indexes are compared within each dataset, the MH D-DIF and D_{STD} indexes generally have similar correlation coefficient patterns. Correlation coefficients for both of these indexes were usually higher than those of χ^2_{MH} and χ^2_L. Figure 1 shows the frequency distribution of correlation coefficients for the four DIF indexes. As noted, most of the correlation coefficients for MH D-DIF were grouped between .50 and .83 and those for D_{STD} were grouped between .60 and .83. For χ^2_L index, most of the correlation coefficients fell between .40 and .58. However, the distribution of correlation coefficients for χ^2_{MH} was scattered and more than half of the twelve within dataset correlation coefficients comparisons were lower than .50. The overall mean for the MH D-DIF coefficients was .59 and for χ^2_{MH} it was .33. The overall mean for the D_{STD} coefficients was .55 and for χ^2_L it was .43. These results indicate that, the MH D-DIF and D_{STD} indexes tend to produce more reliable results in the process of DIF analysis than the χ^2_L. The χ^2_{MH} index is the least stable in the four indexes. These finding are similar to what was found in previously studies (Ryan, 1991; Skaggs & Lissitz, 1992), that the MH D-DIF index had higher correlation coefficients than χ^2_{MH} index. Results from the degree of item classification inconsistency in the latter show that the MH and STD methods have a similar low rate of item classification inconsistency. These results provide strong evidence that the MH D-DIF index was a better indicator than χ^2_{MH} index in establishing the reliability of the MH method.

Insert Figure 1 here

Consistency of Item Classification

Because the criterion to judge item classification in three DIF detection methods were different, and only the LR method distinguished DIF types, it was necessary to make the classification results of these three DIF detection methods comparable. To do this, items with categories of intermediate DIF or large DIF in the MH and STD methods were labeled as "exhibiting DIF". For the LR method, items with categories of non-uniform DIF or uniform DIF were labeled as "exhibiting DIF". Items which were not included in the above categories were labeled as "no DIF" for each method.

Table 7 through Table 9 list the percentages of items labeled as "exhibiting DIF" in each subset based on detection results from the MH, STD, and LR methods. As shown in these tables, the volume of items labeled as "exhibiting DIF" based on the various DIF detection methods was very different. More items were labeled as "exhibiting DIF" in the LR method than the STD and MH methods. The fewest items labeled by the MH method as "exhibiting DIF". The mean percentages of items labeled as "exhibiting DIF" by the MH, STD and LR were 11, 25, and 40%, respectively.

Insert Tables 7, 8, and 9 here

It was also found that items labeled as "exhibiting DIF" by the MH method were also identified as "exhibiting DIF" by the STD and LR methods. Items which were labeled as "exhibiting DIF" by the STD method were also labeled as "exhibiting DIF" by the LR method. For example, in subset A1 of the Spelling test, item #6 which was labeled as "exhibiting DIF" by the MH method was also labeled "exhibiting DIF" by the STD method. The same label was obtained when it was examined by the LR method. It should be noted that, items which were labeled as "exhibiting DIF" by the MH or STD methods could be identified as either uniform DIF or non-uniform DIF in the LR method.

The above results indicate that the LR method is more sensitive than the MH and STD methods. This finding was consistent with those found by Rogers and Swaminathan (1993). However, although the sensitivity of the three DIF detection methods were different, the results from three methods are not mutually exclusive. Items which are identified as "exhibiting DIF" by the MH or STD methods are also found having DIF in the LR method.

The reliability of DIF detection methods was also examined based on the degree of item classification inconsistency between subsets. Table 10 through Table 13 list the item classification inconsistency percentages for each dataset based on detection of MH, STD and LR methods in the four subtests.

Insert Tables 10, 11, 12 and 13 here

As shown in Table 10 through Table 13, the percentages of total item classification inconsistencies for each dataset in the LR method are consistently higher than those in either the MH or the STD methods. The total inconsistency mean percentage for the MH, STD and LR methods were 15, 27 and 39%, respectively. A detailed look at the inconsistencies that happened with the LR method showed that serious inconsistencies were more evident than minor inconsistencies: the mean percentage of serious inconsistencies was 30%, whereas the mean percentage of minor inconsistencies was only 9%. This difference implies that when items were found to have different classifications across two subsets, these classifications tended to show inconsistencies between "no DIF" and "exhibiting DIF" rather than differences in DIF types (i.e., uniform or non-uniform). It appears that detection results from the LR method are not very stable and that item mis-classifications are not just due to flat index values.

Although both the MH and STD methods resulted in low percentages of total inconsistencies for item classifications, the percentages of total inconsistencies with the MH method (the mean percentage = 15%) were typically lower than those with the STD method (the mean percentage = 27%). It was also found that, both of these methods rarely resulted in serious inconsistencies. The mean percentage of serious inconsistencies for the MH method was 3% and for STD method was 2%. This finding implies that even though both of these methods resulted in some inconsistencies in item classifications across two subsets, these inconsistencies were minor and might have resulted from flat index values. Only a few serious inconsistencies of classification from "no DIF" to "large DIF" were found in these two methods.

In summary, both the MH and STD methods tend to identify fewer items exhibiting DIF than does the LR method. This finding suggests that the LR method is more sensitive in detecting

DIF items. However, the characteristic of sensitivity for the LR method does not make its detection more accurate than other two methods. Detection results from both of the MH and STD methods are more stable than the LR method. Fewer items were classified inconsistently across two subsets through the MH and STD methods.

The Influences of Sources of Error Variance on Reliability of DIF Detection Methods

To examine the influence of error variance sources on the reliability of DIF detection methods, correlation coefficients from DIF indexes and the degree of item classification inconsistency from Datasets A, B, and C based on the same DIF detection methods are compared. However, comparisons of correlation coefficients of DIF index and degree of item classification inconsistency exhibited conflicting results for some datasets. For example, for the Spelling test, Dataset B had high correlation coefficients for each MH D-DIF and D_{STD} index which implied MH and STD methods were highly reliable in this dataset, but compared with the other two datasets in the same test, Dataset B also had a high percentage of item classification inconsistencies across subsets which implied it did not produce reliable results. To investigate this conflict, the relationship of DIF indexes between two subsets were plotted. Figure 2 displays the scatter plots of MH D-DIF index between two subsets in each dataset of the Spelling test. The MH D-DIF indexes in subsets B1 and B2 were more scattered than those in the other subsets. Figure 3 is the scatter plots of the D_{STD} indexes between two subsets for each dataset of the Spelling test and it also displays large variability of this index in subsets B1 and B2. It is known that one important factor influencing the size of a correlation coefficient is the nature of the group on which the correlation is measured. Both of the plots suggest that the high correlation of DIF indexes between subsets B1 and B2 could be due to the large variability of index values. Unfortunately, the large variability of index values between subsets also caused more items classified inconsistently.

Insert Figures 2 and 3 here

The opposite happened when the variability of the index values was small. For example, for the Math Computation test, MH D-DIF and D_{STD} indexes were not correlated in Dataset C.

However, compared with the other two datasets, the percentages of item classification inconsistencies based on the MH and STD methods were relatively low in Dataset C. Figure 4 and Figure 5 shows plots of the relationship of MH D-DIF and D_{STD} indexes with themselves between two subsets. These plots indicate that the variability of the MH D-DIF and D_{STD} indexes between subsets C1 and C2 were relatively smaller than those in the other two datasets. It was obvious that the variability of DIF index values played an important role in the correlation analyses and resulted in misleading reliability values. In contrast, the degree of item classification inconsistency across two subsets provided more information when the effects of sources of error variance on reliability were considered. For this reason, only the percentages of item classification inconsistency across two subsets is discussed as an indicator of reliability in this part.

Insert Figures 4 and 5 here

Figure 6 through Figure 9 present the changes of the percentages of item classification inconsistency after three error variance sources were controlled in each dataset. The influence of controlling various error variance on the reliability of DIF detection was visible. Figure 6 displays the changes of percentages on three datasets when the MH method was used in the four subtests. Except for the Reading Comprehension test in which no items were identified as exhibiting DIF, it was found that for the other three tests, Dataset B typically had a higher inconsistency percentage and Dataset C had a lower inconsistency percentage. That is, compared with results of Dataset A, the treatment of controlling the examinee sampling but ignoring the occasion sampling increased the probability of classifying items inconsistently. On the other hand, the treatment of school matching decreases this probability. This tendency was more obvious in the Spelling test. The difference of inconsistencies percentage between Datasets A and B was 33% and that between Datasets A and C was 15%.

Figure 7 presents the percentages of item classification inconsistencies on three datasets when the STD method used in the four subtests. A similar pattern of percentage changes was found. Compared with results from Dataset A, inconsistency percentage increased in Dataset B but slightly decreased in Dataset C. Again, the tendency was more obvious in the Spelling test which showed the difference of inconsistency percentage between Datasets A and B being 19% and that between Datasets A and C being 15%.

The pattern of change in percentage of item classification inconsistencies for the LR method was not as clear as those for the MH and STD methods. As shown in Figure 8, only the Spelling and Math Computation tests increased slightly the inconsistency percentages in Dataset B. Moreover, except for the Reading Comprehension test, three other tests showed decreasing percentages in Dataset C. However, these changes were not salient.

To summarize the above results, the treatment of school matching on Caucasian and African-American students did have some influence on the reliability of DIF detection methods, although this influence varied for different DIF detection methods and for different tests. Item classification inconsistencies across subsets decrease slightly after school matching for both the reference and focal groups. Obvious effects were observed especially in the Spelling test. This finding suggested that different school curricula may play a role in differences of student performance. The DIF detection results are more consistent across subsets when this factor is removed.

Moreover, the results show that controlling examinee sampling in Dataset B did not improve the reliability of DIF detection methods. In contrast, more items were classified inconsistently. A detailed look at Dataset B found that although examinee sampling was controlled, more unexpected factors were included along with the process of treatment. Remember that subset B1 in Dataset B consisted of students who took Level 11 of the *ITBS* test in the fifth grade, and subset B2 included the same students from B1 who later took tests of Level 12 in the sixth grade. Students' cognitive growth from cultural environment and school instruction during this period may change their ability to answer items correctly and interfere with the reliability of DIF detection from this dataset. For example, both Caucasian and African-American students may not able to answer some specific items correctly when they were in the lower grade. However, one group of students may have more opportunities to answer these items correctly based on cultural advantage when they advance to the higher grade. In contrast, some items may exhibit DIF for Caucasian and African-American students in the lower grade. However, differential functioning of items may be eliminated based on school instruction later provide to all students.

Furthermore, different locations of common items in consecutive test levels may also cause the DIF detection results from two subsets of Dataset B to be unreliable. It is known that in the *ITBS* common items are always located in the last part of lower level tests, and the same items are located in the beginning of higher level tests. Differential item functioning may occur only because different group students vary in the rate of speed with which they reach items at the end of a test, or DIF does not happen because both group students are unable to reach items at the end of a test.

Conclusion

The present study compared the reliability of the MH, LR, and the STD methods. Comparisons among the three DIF detection methods found that the MH method usually identified the fewest items as "exhibiting DIF". The LR method tended to label the most items as "exhibiting DIF". However, the apparent sensitivity of the LR method did not make its detection more accurate or stable than the other two methods. The LR index (χ^2_L) usually had lower correlation coefficients and this method produced more item classification inconsistencies across subsets. In contrast, MH D-DIF and D_{STD} indexes had similar high correlation coefficients and both provided a low number of inconsistencies of item classifications across subsets. It implies that both of MH and STD methods produce more reliable and consistent DIF detection results.

The present study also examined the effect of different sources of error variance, namely, examinee, occasion, and curriculum sampling on the reliability of DIF detection. It was found that controlling the error variance due to curriculum sampling decreased slightly the rate of item classification inconsistencies. This finding suggested that different school curriculum may play a role in the differences found in student performance. The reliability of DIF detection is improved when this factor is controlled. However, this study also found that controlling examinee sampling did not improve the reliability of DIF detection and produced somewhat confusing results. The reliability of DIF detection was decreased when larger percentages of item classification inconsistencies happened after the treatment of controlling examinee sampling. Some unexpected factors (such as: student's cognitive growth and location of items in different test administrations) which added along with this treatment may interfere with the true effect. In the future, study on the effects of error variance should consider these factors carefully.

This study also found that more reliability information was provided from the degree of item classification inconsistency than from the correlation analyses of DIF indexes. Since the variability of the DIF index values had obvious influence on the correlation analyses and sometimes resulted in misleading reliability values, the agreement of item classification provided clearer and more direct information about reliability.

Reference

Camilli, G., & Shepard, L.A. (1994). Methods for identifying biased test items. Measurement methods for the social science series, vol. 4. Newbury Park, CA: Sage publications, Inc.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Engelhard, G., Jr., Hansche,L., & Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning. Applied Measurement in Education, 3, 347-360.

Hambleton, R.K., & Jones, R.W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. Educational Research Quarterly, 18(1), 21-36.

Hoover, H.D., & Kolen, M.J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8(2), 173-181.

Plake, B.S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Qualls, A., & Hoover, H.D. (1981, April). Black and white teacher ratings of elementary achievement test items for potential race favoritism. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17(2), 105-116.

Ryan, K.E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. Journal of Educational Measurement, 28(4), 325-337.

Shepard, L.A., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Skaggs, G., & Lissitz, R.W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. Journal of Educational Measurement, 29(3), 227-242.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Table 3: DIF index correlations between two subsets in each dataset for the Reading Comprehension test.

	A	B	C
MH D-DIF	0.364	0.379	0.498*
χ^2_{MH}	0.224	0.016	-.304
D_{STD}	0.283	0.155	0.403*
χ^2_L	0.534*	-.328	0.112

*p < .05

Table 4: DIF index correlations between two subsets in each dataset for the Spelling test.

	A	B	C
MH D-DIF	0.757*	0.830*	0.744*
χ^2_{MH}	0.503*	0.206	0.430
D_{STD}	0.751*	0.753*	0.727*
χ^2_L	0.529*	0.135	0.582*

*p < .05

Table 5: DIF index correlations between two subsets in each dataset for the Usage and Expression test.

	A	B	C
MH D-DIF	0.826*	0.811*	0.502*
χ^2_{MH}	0.311	0.672*	0.616*
D_{STD}	0.826*	0.737*	0.310
χ^2_L	0.528*	0.805*	0.468*

*p < .05

Table 6: DIF index correlations between two subsets in each dataset for the Math Computation test.

	A	B	C
MH D-DIF	0.627*	0.686*	0.278
χ^2_{MH}	0.523*	0.489*	0.300
D_{STD}	0.661*	0.712*	0.230
χ^2_L	0.809*	0.397	0.564*

*p < .05

17

Figure 1: Distributions of correlation coefficients for four DIF indexes.

MH D-DIF

.0	
.1	
.2	8
.3	6 8
.4	
.5	0 0
.6	3 9
.7	4 6
.8	1 3 3
.9	

χ^2_{MH}

.0	0 2
.1	
.2	1 2
.3	0 1
.4	3 9
.5	0 2
.6	2 7
.7	
.8	
.9	

D_{STD}

.0	
.1	6
.2	3 8
.3	1
.4	0
.5	
.6	6
.7	1 3 4 5 5
.8	3
.9	

χ^2_L

.0	0
.1	1 4
.2	
.3	
.4	0 7
.5	3 3 3 6 8
.6	
.7	
.8	1 1
.9	

Table 7: Percents of items labeled as "exhibiting DIF" by the Mantel-Haenszel method in each subset.

	A1	A2	B1	B2	C1	C2
Reading Comprehension	0	0	0	0	0	0
Spelling	5	14	33	33	5	0
Usage & Expression	5	15	15	25	5	5
Math Computation	10	10	20	45	10	5

Table 8: Percents of items labeled as "exhibiting DIF" by the standardization method in each subset.

	A1	A2	B1	B2	C1	C2
Reading Comprehension	0	8	12	28	0	16
Spelling	19	29	43	52	19	29
Usage & Expression	30	35	40	45	20	20
Math Computation	20	20	50	50	15	10

Table 9: Percents of items labeled as "exhibiting DIF" by the logistic regression method in each subset.

	A1	A2	B1	B2	C1	C2
Reading Comprehension	48	36	24	8	20	44
Spelling	57	48	33	38	38	43
Usage & Expression	50	65	45	55	40	30
Math Computation	40	45	30	55	30	30

Table 10: Numbers of item classifications inconsistency in each dataset of the Reading Comprehension subtest [a].

	A			B			C		
	Serious	Minor	Total	Serious	Minor	Total	Serious	Minor	Total
MH	0	0	0	0	0	0	0	0	0
	(0)[b]	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
STD	0	2	2	0	10	10	0	4	4
	(0)	(8)	(8)	(0)	(40)	(40)	(0)	(16)	(16)
LR	9	0	9	8	0	8	10	0	10
	(36)	(0)	(36)	(32)	(0)	(32)	(40)	(0)	(40)

[a] The total number of common items in this subtest is 25.
[b] The value in the parenthesis is the percent of items.

Table 11: Numbers of item classifications inconsistency in each dataset of the Spelling subtest [a].

	A			B			C		
	Serious	Minor	Total	Serious	Minor	Total	Serious	Minor	Total
MH	0	4	4	4	7	11	0	1	1
	(0)[b]	(19)	(19)	(19)	(33)	(52)	(0)	(5)	(5)
STD	0	7	7	2	9	11	0	4	4
	(0)	(33)	(33)	(10)	(43)	(52)	(0)	(19)	(19)
LR	6	3	9	9	1	10	5	2	7
	(29)	(14)	(43)	(43)	(5)	(48)	(24)	(10)	(33)

[a] The total number of common items in this subtest is 21.
[b] The value in the parenthesis is the percent of items.

Table 12: Numbers of item classifications inconsistency in each dataset of the Usage and Expression subtest [a].

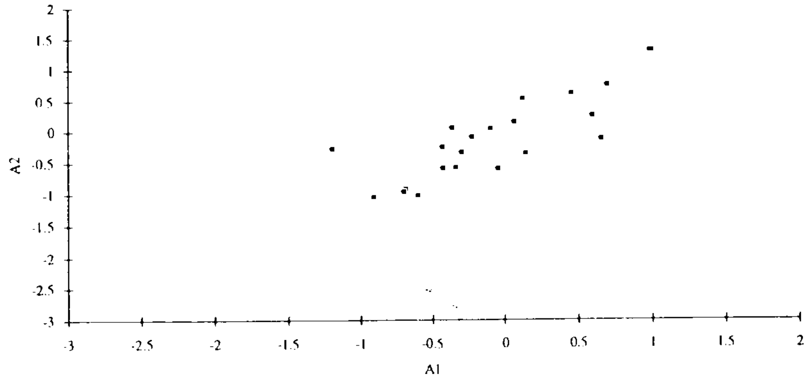
	A			B			C		
	Serious	Minor	Total	Serious	Minor	Total	Serious	Minor	Total
MH	0	3	3	1	3	4	0	2	2
	(0)[b]	(15)	(15)	(5)	(15)	(20)	(0)	(10)	(10)
STD	0	5	5	0	8	8	0	4	4
	(0)	(25)	(25)	(0)	(40)	(40)	(0)	(20)	(20)
LR	5	5	10	4	5	9	6	1	7
	(25)	(25)	(50)	(20)	(25)	(45)	(30)	(5)	(35)

[a] The total number of common items in this subtest is 20.
[b] The value in the parenthesis is the percent of items.

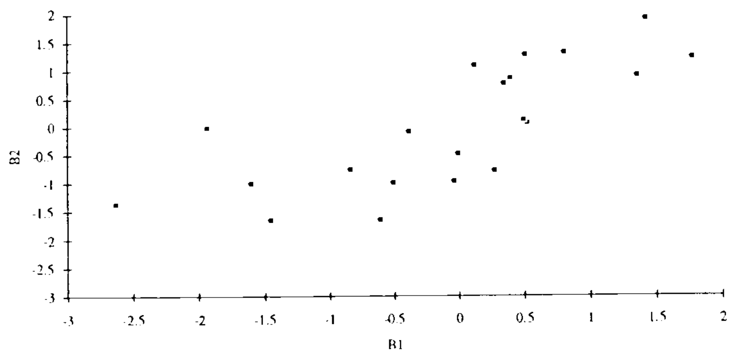
Table 13: Numbers of item classifications inconsistency in each dataset of the Math Computation subtest [a].

	A			B			C		
	Serious	Minor	Total	Serious	Minor	Total	Serious	Minor	Total
MH	0	3	3	3	4	7	0	1	1
	(0)[b]	(15)	(15)	(15)	(20)	(35)	(0)	(5)	(5)
STD	0	5	5	1	5	6	0	3	3
	(0)	(25)	(25)	(5)	(25)	(30)	(0)	(15)	(15)
LR	5	1	6	7	2	9	4	1	5
	(25)	(5)	(30)	(35)	(10)	(45)	(20)	(5)	(25)

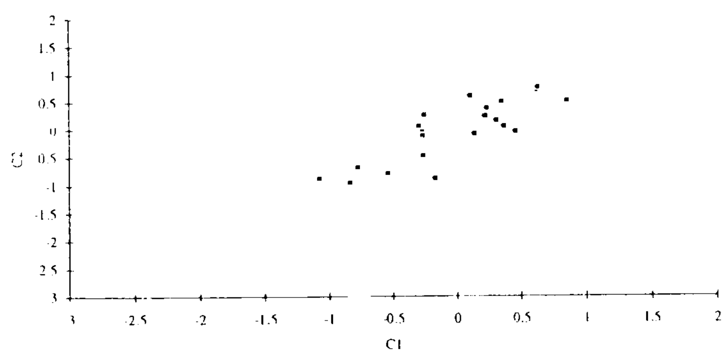
[a] The total number of common items in this subtest is 20.
[b] The value in the parenthesis is the percent of items.



2(A): Scatter plot between A1 and A2 (r = .76).



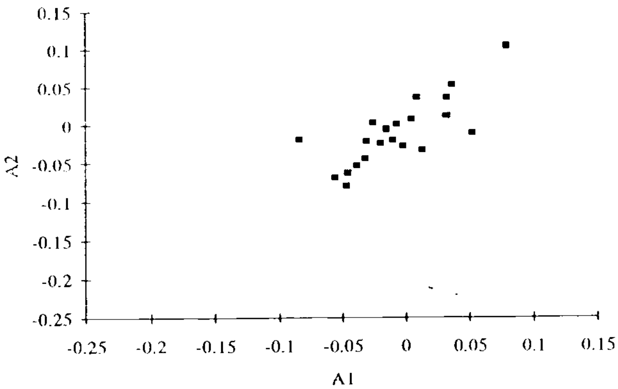
2(B): Scatter plot between B1 and B2 (r = .83).



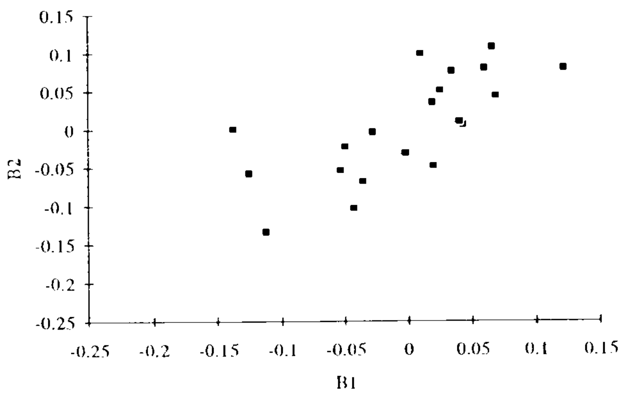
2(C): Scatter plot between C1 and C2 (r = .74).

Figure 2: The Variability of MH D-DIF Indexes between Two Subsets in Each Dataset for the Spelling Test.

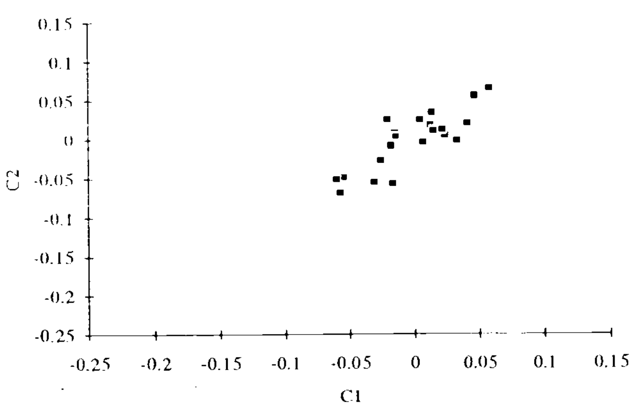
21



3(A): Scatter plot between A1 and A2 (r = .75).

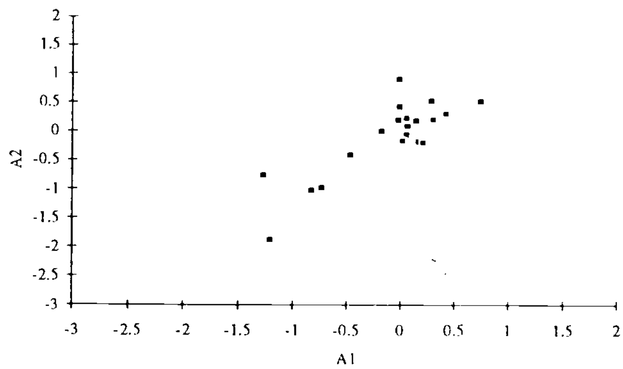


3(B): Scatter plot between B1 and B2 (r = .75).

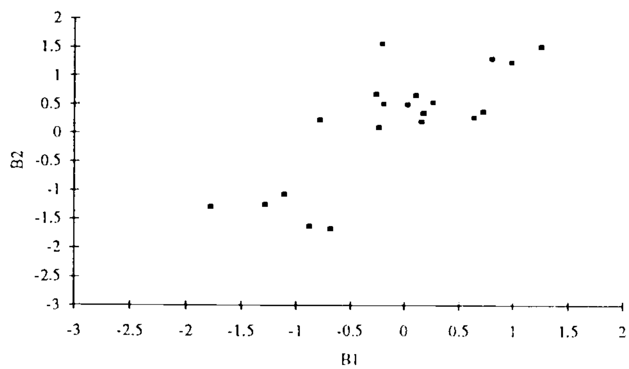


3(C): Scatter plot between C1 and C2 (r = .73).

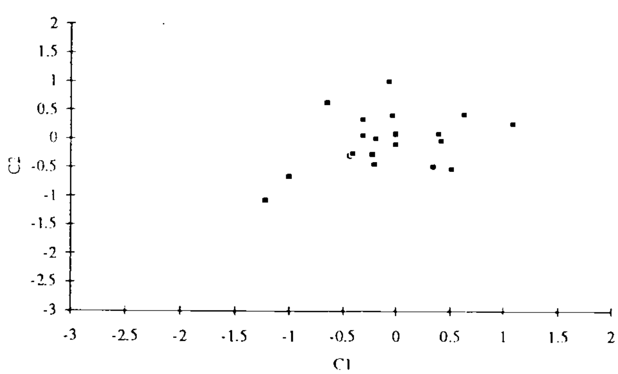
Figure 3: The Variability of D_{STD} Indexes between Two Subsets in Each Dataset for the Spelling Test.



4(A): Scatter plot between A1 and A2 (r = .63).



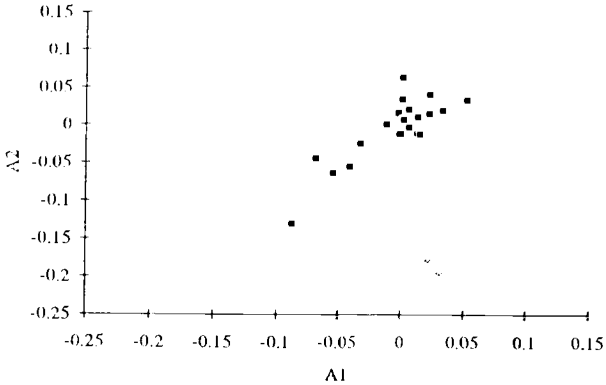
4(B): Scatter plot between 1B and 2B (r = .69).



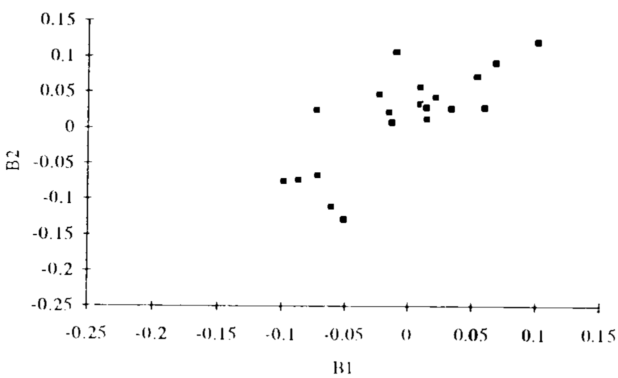
4(C): Scatter plot between C1 and C2 (r = .28).

Figure 4: The Variability of MH D-DIF Index between Two Subsets in Each Dataset for the Math Computation Test.

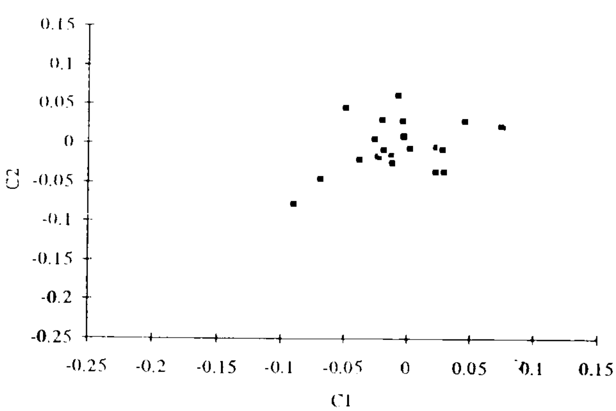
23



5(A): Scatter plot between A1 and A2 (r = .66).



5(B): Scatter plot between B1 and B2 (r = .71).



5(C): Scatter plot between 1C and 2C (r = .23).

Figure 5: The Variability of D_{STD} Index between Two Subsets in Each Dataset for the Math Computation Test.

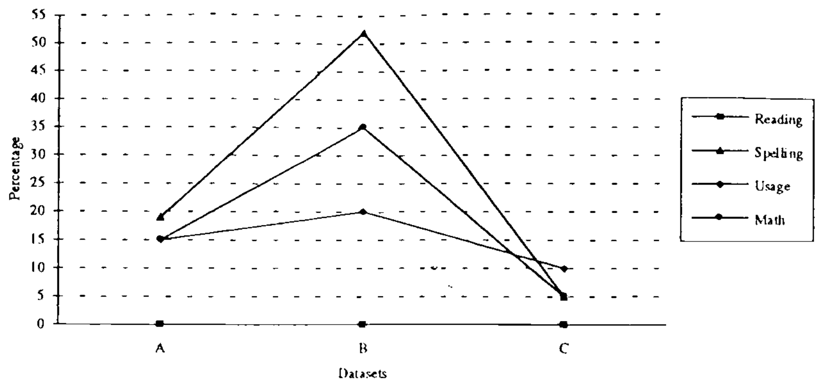


Figure 6: Percentages of item classification inconsistencies on the Mantel-Haenszel method.

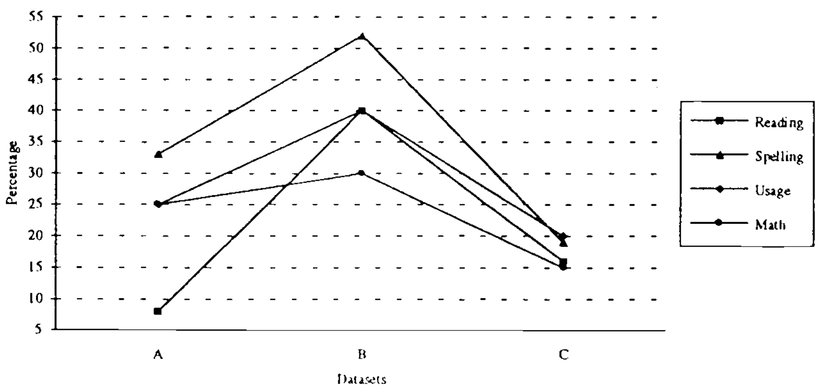


Figure 7: Percentages of item classification inconsistencies on the standardization method.

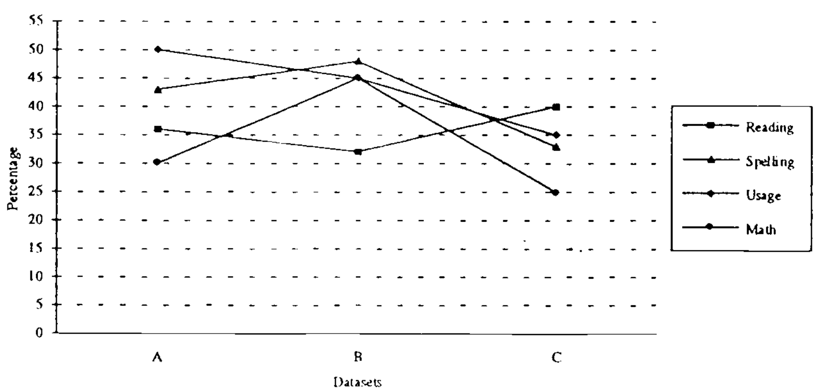


Figure 8: Percentages of item classification inconsistencies on the logistic regression method.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028408

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Factors Influencing the Reliability of DIF Detection Methods

Author(s): Chi-Yu Huang & Stephen B. Dunbar

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Chi-Yu Huang	Printed Name/Position/Title: Chi-Yu Huang / Psychometrician	
Organization/Address: ACT, Inc. 2201 N Dodge Street, Iowa City, IA 52243	Telephone: (319) 337-1582	FAX:
	E-Mail Address: huang@act.org	Date: 4/17/'98

(over)

Stephen B. Dunbar / Professor   The University of Iowa   (319) 335-5561

Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 20, 1998

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae.net.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (424)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1998/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an AERA chair or discussant, please save this form for future use.

The Catholic University of America