DOCUMENT RESUME

ED 193 265 TM 800 580

AUTHOR Harnisch, Delwyn L.; Linn, Robert L.
TITLE Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices.
PUB DATE Apr 80
NOTE 27p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Curriculum; Elementary Secondary Education; Response Style (Tests); Social Influences; Student Motivation; Test Anxiety; Test Bias; *Testing Problems; *Test Items; *Test Validity
IDENTIFIERS *Test Appropriateness; *Test Content

ABSTRACT

Indices of appropriateness of a test for an individual are discussed, and two data sets are evaluated. With the first data set, three indices of test appropriateness are obtained for response patterns on achievement tests from an experimental study of the effects of test anxiety and time pressure with 173 3rd and 4th grade students. Relationships of these three indices to student background characteristics and measures of anxiety are of special interest in this first data set. With the second data set, two indices of test appropriateness are obtained for the math and reading test given to 6300 students from a random sample of approximately 110 schools at 4th, 8th, and 11th grade levels. Relationships of these indices to student sociocultural background and measures of student anxiety and motivation are examined at the individual student level and school level. The test appropriateness measures have potential utility in the identification of students for whom the test is inappropriate or schools with curricula not matching test content. (Author/GK)

Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices

Delwyn L. Harnisch and Robert L. Linn

University of Illinois, Urbana-Champaign

Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April, 1980

Abstract

This research is concerned with appropriateness of tests. Two data sets are evaluated in this study. With the first data set, three indices of test appropriateness are obtained for response patterns on achievement tests from an experimental study of the effects of test anxiety and time pressure with 173 3rd and 4th grade students. Relationships of these three indices to student background characteristics and measures of anxiety are of special interest in this first data set. With the second data set, two indices of test appropriateness are obtained for the math and reading test given to 6300 students from a random sample of approximately 110 schools at 4th, 8th, and 11th grade levels. Relationships of these indices to student socio-cultural background and measures of student anxiety and motivation are examined at the individual student level and school level. The test appropriateness measures have potential utility in the identification of students for whom the test is inappropriate or schools with curricula not matching test content.

Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices

Traditional test theory and interpretations of test results are based on an implicit assumption that the test is equally appropriate and reliable for all examinees. Additionally, it is assumed that the test results are adequately summarized by a single global score. In general, test fallibility is recognized and considerable emphasis is given to coefficients of reliability and to errors of measurement. However, the errors of measurement are assumed to be non-systematic, i.e., to be unrelated to any student characteristics. Furthermore, the variance of the errors of measurement is assumed to be constant across all levels of ability.

The assumption that a test is equally appropriate for all individuals is a strong and important one. In this paper, it is our objective to investigate the tenability of this assumption. It is intuitively reasonable that a standardized test may be more appropriate for some individuals than others. Unique background experiences may make an item very difficult that is quite easy for most people, i.e., the child who has never gone camping may find a reading passage about a camping experience more difficult than do children who have had the experience. Individual differences in motivational dispositions such as test anxiety may make an item very difficult that is quite easy for most people. Differential exposure to and emphasis of subject matter covered by an achievement test may result not only in mean differences on the total score from class to class but in differences in typical response patterns. Subtests of test items that are relatively difficult for most students may be relatively easy for students who have been in classes where that particular content was emphasized. Such aberations may lead to the systematic over- or under-estimation of an individual's or group's level of achievement. That is, they tend to distort the measurement results.

Indices of appropriateness of a test for an individual could be used in a variety of ways. They could lead to the identification of subgroups of people for whom the test is inappropriate. In addition, the items that contribute most to an index indicating that the test is inappropriate for particular subgroups might also be identified. Furthermore, the content of such items could be analyzed toward the eventual end of reducing inappropriateness through test revision.

There are two major types of appropriateness indices based upon the pattern of responses to individual items. First, there are the indices which are based upon latent trait models as described by Levine and Rubin (1976) and modifications of these indices as suggested by Drasgow (1978). Second, there are the indices which are based directly upon the pattern of right and wrong answers, such as the "caution" index proposed by Sato (1975) and the "U" index by Henk van der Flier (1977). The latter type of appropriateness indices based upon the pattern of responses to individual items will be investigated further in this paper.

The investigation of response patterns starts with a matrix of ones and zeros. A row of the matrix is associated with each examinee and a column with each item. Ones are recorded for correct responses and zeros for incorrect responses. Rows and columns of the data matrix are permuted so that the items (columns) are arranged, from left to right in ascending order of difficulty, and examinees (rows) are arranged, from top to bottom in descending order of total number of correct answers. The resulting matrix has been called an "S-P Table" (student-problem table) by Sato (1975; see Tatsuoka, Note 2, for a description in English).

If the items formed a perfect Guttman Scale (Guttman, 1941) the S-P Table would consist of a section of all ones in the upper left-hand corner and all

zeros in the lower right-hand corner. A single step-like boundary line would separate the ones and the zeros. In other words, anyone who responded correctly to a difficult item would also answer all easier items correctly. Of course, with responses to achievement test items perfect Guttman scales cannot be expected. Consequently, the S-P Table will be characterized by a predominance of ones in the upper-left hand corner and zeros in the lower right-hand corner, but there will be many exceptions to the pattern, i.e., ones in the region where mostly zeros are found and vice versa.

A small hypothetical example of an S-P Table with 18 examinees and 5 items is shown in Table 1. The solid and dashed lines in Table 1 are known as the S-curve and P-curve respectively. The S curve (solid line) is obtained by drawing a vertical line for each row that has $n_{i.}$ items (columns) to the left of it where $n_{i.}$ is the total number of correct responses for the $i^{\underline{th}}$ examinee. The P curve (dashed line) is obtained by drawing a horizontal line in each column such that there are $n_{.j}$ examinees (rows) above it, where $n_{.j}$ is the number of examinees who answer item j correctly.

For an ideal, or Guttman-scalable S-P Table, the S and P curves would coincide. The degree of divergence provides an indication of the degree of homogeneity of the response patterns. Sato (1975) has developed an index based on the area between the S and P curves which is potentially useful in evaluating the homogeneity of the test (see Tatsuoka, Note 2). Of greater interest for purposes of this paper, however, is Sato's "caution" index.

Sato's caution index, C_i for the $i^{\underline{th}}$ examinee, may be defined as follows:

$$C_i = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{ij})\, n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij}\, n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - n_{i.} \left(\frac{\sum_{j=1}^{J} n_{.j}}{J} \right)}$$

where $i = 1,2,\ldots I$, indexes the examinee,

$j = 1,2,\ldots J$, indexes the item,

$u_{ij} = \begin{cases} 1 & \text{if examinee } i \text{ answers item } j \text{ correctly,} \\ 0 & \text{if examinee } i \text{ answers item } j \text{ incorrectly} \end{cases}$

$n_{i.}$ = total correct for the $i^{\underline{th}}$ examinee, and

$n_{.j}$ = total number of correct responses to the $j^{\underline{th}}$ item.

A parallel index for the $j^{\underline{th}}$ item may be defined by simply reversing the roles of i and j in the above equation, but only the person index will be considered in the present paper.

The name of the index comes from the notion that a large value is associated with examinees (or items) that have unusual response patterns. It denotes that some caution may be needed in interpreting a total correct score for an examinee (item). An unusual response pattern may result from carelessness, from high student anxiety, from an unusual instructional history or other experiential background, from a localized misunderstanding that influences responses to a subset of items, or possibly from copying a neighbor's answers to certain questions. The key point is that the caution index provides information about an examinee (or item) that is not contained in the total score. A large value of the caution index raises doubts about the validity of the usual interpretation of the total score for an individual.

Table 1

S-P Table for 18 Examinees and 5 Items
(Hypothetical Example)

Examinee i	Item j 1	2	3	4	5	Examinee Total n_i	Sato's Caution Index C_i	Modified Caution Index C_i^*	van der Flier's Index U_i'
1	1	1	1	1	0	4	.00	.00	.00
2	1	1	1	0	1	4	.65	.33	.25
3	1	1	1	0	0	3	.00	.00	.00
4	1	1	0	1	0	3	.16	.08	.17
5	1	1	0	0	1	3	.65	.31	.33
6	1	0	1	0	1	3	1.13	.54	.50
7	1	1	0	0	0	2	.00	.00	.00
8	1	1	0	0	0	2	.00	.00	.00
9	1	0	1	0	0	2	.44	.23	.17
10	1	0	0	1	0	2	.59	.31	.33
11	0	1	1	0	0	2	.74	.39	.33
12	0	1	0	1	0	2	.88	.47	.50
13	1	0	0	0	0	1	.00	.00	.00
14	1	0	0	0	0	1	.00	.00	.00
15	0	1	0	0	0	1	.45	.22	.25
16	0	0	1	0	0	1	1.14	.56	.50
17	0	0	0	1	0	1	1.36	.67	.75
18	0	0	0	1	0	1	1.36	.67	.75
Item Total $n_{.j}$	12	10	7	6	3				
Sato's Caution Index C_j	.30	.28	.42	.95	.21				
Modified Caution Index C_j^*	.14	.14	.21	.50	.13				
van der Flier's Index U_j'	.06	.16	.21	.57	.16				

A modified form of Sato's caution index, C_i^*, was introduced to yield a lower bound of 0 and an upper bound of 1. This modified caution index, C_i^*, for the $i^{\underline{th}}$ examinee may be defined as follows:

$$C_i^* = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{ij}) n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij}\, n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=J+1-n_{i.}}^{J} n_{.j}}$$

Similarly, a parallel index for the $j^{\underline{th}}$ item on the modified caution index may be defined by simply reversing the roles of i and j in the above equation. The resulting values of C_i, C_i^*, C_j, and C_j^* are computed and shown in Table 1.

The third appropriateness index used in our investigation was proposed by Henk van der Flier (1977). Using the order of the items and the subjects as in the S-P table, the deviation from the S-curve can be quantified by adding the number of 1's to the right of every 0 which is called U (it resembles the Mann Whitney U). The minimum value of U is 0 and the maximum value equals the number of correct answers multiplied by the number of incorrect answers. The range of U values are made equal for different total scores by dividing them by this maximum value. The resulting measure (U_i') has a lower bound of 0 and an upper bound of 1. A parallel index for the $j^{\underline{th}}$ item, U_j', although not mentioned by Henk van der Flier, may be defined by simply computing the deviation from the P curve by adding the number of 1's to the bottom of every 0. The U_i' and U_j' values are computed and shown in Table 1.

Our general interest in studying the appropriateness indices outlined above is to understand the properties and the interrelationships of the indices to

each other and to other variables. Student background characteristics and measures of anxiety are of special interest in the first set of analyses to be reported.

Results and Discussion on Anxiety Study

The first data set we used for analyses of information contained in response patterns on achievement tests comes from an experimental study of the effects of test anxiety and time pressure on achievement test performance (Plass & Hill, Note 1). Low, middle, and high anxious students were randomly assigned to take tests either with or without time pressure. Within each of these conditions a second test was administered to randomly selected subgroups with standard instructions, work faster instructions, and work slower instructions. A group of 173 3rd and 4th grade students were involved in this study.

The indices were obtained for all students under each of the time pressure conditions. Students with a zero or perfect score were not included in the correlational study since no new information is gained from a response pattern yielding a zero or perfect score. Both groups received arithmetic word problem tests, 30 items in length, where the problems selected for the test were similar both to those which the children worked in the classroom and to problems found on achievement tests. The tests at time two contained essentially the same algorithms as the tests at time one. For example, the number "7" was substituted for a "5", or a "26" was substituted for a "32" denoting that the numbers were changed while the general magnitude of the numbers used remained the same. Similarly, the "story" of each word problem was changed, for example, substituting "x apples and y oranges" for "w boys and z girls". Following their test taking period the children responded to four multiple choice survey items. Two of the four questions have been selected for inclusion. The first question

asked the children to rate the difficulty of the problems while the second question asked them to rate how well they thought they did on the problems.

The indices were then correlated with sex of the subject, test anxiety score, test score, percent of items correct of those attempted, time spent per problem, rating of the difficulty of problem, and rating of their performance. These correlations were computed separately for students in each time pressure condition at both times of testing. These correlations are shown in Table 2.

Significant correlations were found between all the indices and test score, percent of items correct, and time per problem at both times of testing and for students in each time pressure condition. Consistently, stronger correlations exist between test score and each of the indices for students taking the test under the no time pressure condition. This suggests that low ability students are attempting more problems on the test and possibly arriving at a correct response to an item that is normally not answered correctly by a student of their ability. That is, when students are given more time to take a test more items are attempted with the likelihood for low ability students of getting a correct response to items normally not attempted which results in a high caution score.

The relationship of test anxiety score to each of the indices is positive at both times of testing indicating a somewhat greater likelihood of aberrant response patterns for high anxious children than for low anxious children. Furthermore, these correlations, significant at test-time two, suggests that the high anxious students as indicated by their score on the Test Anxiety Scale for Children (Feld & Lewis, 1969) have a more unusual response pattern at test time two than they had at test time one.

Since the caution indices and test anxiety are negatively correlated with total test score, an alternative interpretation of the positive relationship

between anxiety and the caution indices is that they merely reflect a confounding of the caution indices with total score in this data set coupled with the well established negative correlation between test anxiety and total score. Because of the negative correlation between the indices with total score, partial correlation coefficients were computed controlling for the performance of the children on the test as is shown in Table 3.

As can be seen in Table 3, partialling out total score does reduce the magnitude of the correlations between anxiety and the caution indices. Some of the other partial correlations in Table 3 are worthy of note, however. The negative partial correlations of time per problem with the indices suggests that the student hurriedly taking the test results in a students having a rather high caution score alerting us to use caution in interpreting the total test score. Even more suggestive are the differences between the partial correlation between the indices and percent correct under the two time pressure conditions. These partial correlations are negative and generally significant when the test is administered under the time pressure condition but are essentially zero for the no time pressure condition. Time pressure may increase the number of careless responses of students who hurry too fast thus increasing their caution scores while reducing their percent correct scores.

The test retest correlation among the indices are shown in Table 4. The results indicate the strong relationship among all three indices. All correlations between pairs of indices are .96 or higher. The test-retest correlations for the students under no time pressure show higher correlations (.50 to .53) than the students under time pressure (.31 to .42).

In summary, the results from the anxiety study with the indices showed a positive relationship with the test anxiety score when not controlling for the students total test score, but not much relationship when total score is

partialled out. The student hurriedly taking the test under both conditions of time pressure results in him or her having high caution scores. The strong negative correlations (zero order and partial) alerts us to a possible overestimate of their ability. The test-retest correlations indicate that unusual response patterns are more consistent for students under the no time pressure condition than for students under the time pressure condition. Evaluation of the scattergrams of the indices with the total test scores reveals the overestimation of the caution index by both van der Flier's and Sato's caution index for students with very low scores who get correct an item with a low p values. Thus, the authors prefer the modified index over the other two indices because of its property of not overestimating the caution score in the above mentioned case. For further investigation as noted in this paper with our second data set, we will consider the Sato caution index along with the modified caution index.

Results and Discussion from the Fourth Grade Statewide Test

The second data set we used to investigate the utility of the caution indices was the test results from the 1978 Illinois statewide assessment program. This annual statewide survey, known as the Illinois Inventory of Educational Progress (IIEP), provides data for some 6300 students from a random sample of approximately 110 schools at each of the 4th, 8th, and 11th grade levels. At each of the attendance centers, between 18-22 students are randomly selected for the grade level being tested at the school. This gives an approximate statewide sample of 2100 students per grade level (4, 8, and 11). The total IIEP state survey sampling, then, involves some 6300 students spread evenly across the elementary (4th), junior high (8th), and high school (11th) grades being tested.

The data set includes item response data for tests of reading and mathematics. Also included are questionnaire responses which provide data on student background, self-report of test anxiety, their attributions for success and failure, and other specific and general motivation variables.

In the time available, it is impossible to report on the results of the indices for all tests at all three grade levels included in the IIEP. Instead, we have decided to focus on one grade level, fourth, and the tests of reading and mathematics. Thus, the results may be taken as illustrative of the type that may be obtained at other grade levels and on other tests.

The fourth grade mathematics test contains 40 items while the fourth grade reading test contains 28 items. Both caution indices were computed on each test for the 2094 students at the fourth grade.

Our primary focus with the second data set has been on possible school and regional differences in caution indices. As a first step, we used an hierarchical analysis of variance (ANOVA) to disentangle the variance attributable to regions, schools which were nested in regions, and students nested within schools within regions.

These two indices for each test comprise the four dependent measures in a hierarchical ANOVA with school being our unit of analysis. The first factor is the five different regions of the state. School is the second factor which is nested in regions while students are nested within schools within regions. There are 50, 14, 18, 15, and 13 schools in regions one through 5 respectively. Finn's Multivariance program (1977) was used to analyze school effects within region and region effects. The results of these analyses are summarized in Table 5. Significant multivariate region effects were found with this being mainly attributable to the Sato caution index on the reading test. The schools within regions 1, 2, and 5 have a significant multivariate effect with it being

attributed to all the indices in regions 1 and 5 while only being attributed to the math indices for schools within region 2.

The mean caution indices on both tests for the 13 schools within region 5 are reported in Table 6. The range of Sato's caution index is .37 to .63 on the math and .23 to .48 on the reading test. The range of the modified caution index is .18 to .32 on the math test and .13 to .24 on the reading test. The relatively wide range of caution indices for schools within region 5 suggests a high degree of variability of item response patterns in schools. Additionally, these large differences on the caution indices found for schools within region 5 may well have been a function of the curriculum offerings. The school effects are strong for all indices revealing the fact that certain schools possibly have not covered segments of the content being sampled on the test or have just given different emphasis to some aspects of the content than is typical.

The significant schools within region effects denotes the high degree of variability of student performance in schools within certain regions of the state. Curriculum offerings may very well contribute to these large differences found for certain schools within regions. To explore this possibility we did a more detailed analysis of the response patterns of students at schools with high means on the caution indices to identify the subset of items that are contributing most to the caution indices for the identified schools. The analysis of these subsets of items were used to describe unique patterns of performance by item content which suggests differences in content coverage thus making the test less appropriate for some schools than others.

Results and Discussion of Schools with Large Modified Caution Index

Schools with high mean modified caution indices on the math test, .30 or greater, were evaluated for unusual response patterns. The p values were computed for each school on each of the 40 items. Since school mean performance

on the test is directly related to the p values on the items, a linear regression was performed on the p values for each school with the p values from the state sample. The regression equation was used to compute the expected proportion correct on each item for each school. Residual scores were simply computed by subtraction of expected from observed proportion correct on each item for each school.

In attempting to find clues as to the possible reasons for the large differences in the residuals we categorized items based on their content and format. The mean of the residuals was then computed within each content category for each school. The mean residuals were then standardized by dividing by the standard error of estimate. Finally, the standardized mean residuals were multiplied by the square root of the number of items in the content category as a means of weighing the standardized mean residuals according to the number of items in the category. The resulting weighted standardized mean residuals, which are analogous to a critical ratio, which were used to compare the items in different categories, are reported in Table 7.

An entry in Table 7 greater than 2.0 in absolute values indicate that items in that particular category are much easier or much harder for students in that school than would be expected from their overall performance and the relative difficulty of these items for the statewide sample as a whole. Four of the content areas have entries in Table 7 greater than 2.0 in absolute value for one or more schools. These categories are items using figures to represent fractions, story problems dealing with money*, numeration questions, and items involving the metric system. The large negative entry (-2.3) for school 1 for

*An example of a story problem dealing with money is: "Mary earned $1.00 raking leaves. Candy bars cost 15¢. How many candy bars can she buy with her money?"

the figural representation of fractions stands in marked contrast to the large positive values for schools 2, 3, and 7 in this category (4.2, 2.9, and 4.3 respectively). This suggests the hypothesis that the use of figures to represent fractions may be quite common in schools 2, 3, and 7 but rare in school 1.

Similar hypotheses are suggested by the other large values in Table 7. Thus, one would expect that special emphasis is placed on the metric system in schools 6 and 7.

Differences such as those reported above indicate, at the first level, the type of items that function differently for different schools. However, it leaves unanswered the more interesting question of why. We have some hunches that we are presently pursuing, but much additional work is needed to provide any support for them. Differences in response patterns, for example, could result from attendance patterns and school to school variability in content coverage and emphasis. This latter possibility is currently being pursued in this years' statewide testing program with teachers being asked to respond to such questions for each item as: (1) When were students exposed to the item content? (a) have never been (b) prior to this grade level (c) during this grade level; (2) To what extent have students been exposed to the item content? (a) have never been (b) hardly (c) somewhat (d) quite (e) very much; and (3) What percentage of students will answer this item correctly? Information gathered from the teachers on the above questions is intended to give us direct input to the question of content coverage, emphasis, and accuracy. The comparison of observed and expected performance of students attending specific schools as indentified by the above analysis would then be compared along the lines used for the results that we have just presented. With such an analysis we expect that what may appear now as items or categories

being covered thoroughly would be validated by the teacher response data and similarly for the items or categories reported as not being covered thoroughly.

Conclusion

The appropriateness indices of a test for an individual could be used in a variety of ways. The measures of test appropriateness may identify students for whom the test is inappropriate or schools with curricula that do not match the test content. Sato (1975) has also suggested that the caution index may be used along with the total test score to identify students who need more study, who make careless mistakes, who possess sporadic study habits or insufficient readiness or who are doing everything fine. Schools can similarly be identified as having in general more of one of the above type of students. The analyses may also be used to identify items that are part of tests which are providing a misleadingly low indication of the capabilities of minority and/or socioeconomically disadvantaged children as well as the anxious, low performing children of all backgrounds. Furthermore, the content of such items could be analyzed toward the eventual end of reducing inappropriateness through test revision.

Reference Notes

1. Plass, J. & Hill, K. T. Optimizing children's achievement test performance: The role of time pressure, evaluative anxiety and sex. Paper presented at the Annual Meetings of the Society for Research in Child Development, San Francisco, March, 1979.

2. Tatsuoka, M. M. Recent psychometric developments in Japan: Engineers grapple with educational measurement problems. Paper presented at the ONR Contractors Meeting on Individualized Measurement, Columbia, Missouri, 1978.

References

Drasgow, F. Statistical indices of the appropriateness of aptitude test scores. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, 1978.

Feld, S. C., & Lewis, J. The assessment of achievement anxieties in children. Achievement-Related Motives in Children, 1969, 151-199.

Finn, J. D. MULTIVARIANCE: Univariate and multivariate analysis of variance, covariance, regression and repeated measures. Department of Educational Psychology, State University of New York at Buffalo, September, 1977.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, et al. (Eds.), The prediction of personal adjustment. NY: Social Science Research Council, 1941.

Levine, M. V., & Rubin, D. B. Measuring appropriateness of multiple-choice test scores. (Research Bulletin 76-31). Princeton, NJ: Educational Testing Service, 1976.

Sato, T. The construction and interpretation of S-P tables. Tokyo: Meiji Tosho, 1975. (In Japanese).

van der Flier, H. Environmental factors and deviant response patterns. In Y.H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets and Seitlinger, B.V., 1977.

Table 2

Correlations Between Caution Indices and Selected Variables

For Two Groups[a] Based on Time Pressure Condition[b]

At Two Times of Testing

Variables	Grades 3 - 4					
	Sato's Caution Index		Modified Caution Index		van der Flier's Index	
Test Time 1						
Sex[c]	-.17	(-.18)	-.16	(-.17)	-.17	(-.19)
Test Anxiety Score	.09	(.10)	.08	(.08)	.09	(.13)
Test Score	-.27**	(-.36)**	-.24*	(-.35)**	-.31**	(-.41)**
Percent of Items Correct	-.33**	(-.36)**	-.31**	(-.36)**	-.36**	(-.41)**
Time Per Problem	-.30**	(-.23)*	-.30**	(-.21)	-.27**	(-.24)*
Task Difficulty	-.07	(.08)	-.06	(.10)	-.09	(.08)
Performance	-.14	(-.06)	-.14	(-.06)	-.13	(-.10)
Test Time 2						
Sex	-.10	(-.32)**	-.11	(-.36)**	-.11	(-.29)**
Test Anxiety Score	.23*	(.22)*	.19*	(.15)	.21*	(.22)
Test Score	-.39**	(-.57)**	-.24*	(-.40)**	-.43**	(-.63)**
Percent of Items Correct	-.47**	(-.57)**	-.35**	(-.40)**	-.51**	(-.60)**
Time Per Problem	-.25*	(-.50)**	-.23*	(-.41)**	-.26*	(-.49)**
Task Difficulty	.12	(.12)	.11	(.05)	.12	(.15)
Performance	-.04	(-.25)**	.04	(-.13)	-.07	(-.28)**

[a]The sample size for the time pressure condition is 84 at test time 1 and 77 at test time 2 while the sample size for the no time pressure condition is 87 at test time 1 and 71 at test time 2.

[b]Correlations for the no time pressure condition in parenthesis.

[c]Sex coded as one for males.

*p < .05

**p < .01

Table 3

Partial Correlations[a] Between Caution Indices and Selected Variables for Two Groups[b] Based On Time Pressure Condition[c] at Two Times of Testing

Variables	Grades 3 - 4					
	Sato's Caution Index		Modified Caution Index		van der Flier's Index	
Test Time 1						
Sex[d]	-.13	(-.13)	-.12	(-.11)	-.13	(-.13)
Test Anxiety Score	.00	(.03)	.00	(.01)	-.01	(.05)
Percent of Items Correct	-.23*	(-.05)	-.22	(-.04)	-.19	(-.03)
Time Per Problem	-.35**	(-.16)	-.34**	(-.14)	-.32**	(-.16)
Task Difficulty	-.13	(.02)	-.12	(.04)	-.16	(.01)
Performance	-.03	(.16)	-.04	(.15)	-.01	(.15)
Test Time 2						
Sex	-.07	(-.38)**	-.09	(-.39)**	-.08	(-.35)**
Test Anxiety Score	.14	(.04)	.13	(.03)	.10	(.04)
Percent of Items Correct	-.31*	(.01)	-.30*	(.01)	-.31*	(-.01)
Time Per Problem	-.21	(-.27)*	-.21	(-.25)*	-.23*	(-.24)*
Task Difficulty	.06	(-.10)	.07	(-.10)	.04	(-.09)
Performance	.17	(.10)	.17	(.12)	.15	(.08)

[a]Controlled for test Score.

[b]The sample size for the time pressure condition is 84 at test time 1 and 77 at test time 2 while the sample size for the no time pressure condition is 87 at test time 1 and 71 at test time 2.

[c]Partial correlations for the no time pressure condition in parenthesis.

[d]Sex coded as one for males.

*p < .05

**p < .01

Table 4

Correlations[a] Among Caution Indices For Two Groups[b]

Based on Time Pressure Condition at Two Times of Testing[c]

	Sato's Caution Index (I)	Modified Caution Index (I)	van der Flier's Index (I)	Sato's Caution Index (II)	Modified Caution Index (II)	van der Flier's Index (II)
Sato's Caution Index (I)	1.00	.99	.99	.35	.32	.39
Modified Caution Index (I)	.99	1.00	.98	.33	.31	.38
van der Flier's Index (I)	.99	.99	1.00	.38	.35	.42
Sato's Caution Index (II)	.53	.52	.56	1.00	.99	.98
Modified Caution Index (II)	.51	.50	.53	.97	1.00	.96
van der Flier's Index (II)	.50	.49	.53	.98	.96	1.00

[a]Correlations for the time pressure group are in the upper triangle while the correlations for the no time pressure group are in the lower triangle.

[b]The sample size for the time pressure condition is 84 at test time 1 and 77 at test time 2 while the sample size for the no time pressure condition is 87 at test time 1 and 71 at test time 2.

[c]The Roman numeral I represents the index at test time 1 and the Roman numeral II represents the index at test time 2.

Table 5

Hierarchal Analyses of Variance for

Caution Indices on Two Tests

			Schools Within Region				
	Region	Schools	1	2	3	4	5
Multivariate F	5.35**	2.74**	4.05**	2.04**	1.17	1.12	2.27**
Univariate F							
Sato's Caution Index-Math	.67	1.85**	2.12**	2.37*	.45	1.57	2.46**
Modified Caution Index-Math	.57	1.64**	1.69*	2.78**	.46	1.37	2.17*
Sato's Caution Index-Reading	5.65*	2.49**	3.59**	1.43	.84	1.12	3.12**
Modified Caution Index-Reading	2.12	1.74**	1.97**	1.73	.66	1.35	2.83**

* $p < .01$

** $p < .001$

Table 6

Mean Caution Indices on Two Tests

For Schools Within Region 5

		Math Test		Reading Test	
School Number	Sample Size	Sato's Caution Index	Modified Caution Index	Sato's Caution Index	Modified Caution Index
1	20	.44	.23	.25	.15
2	20	.55	.27	.41	.23
3	11	.51	.26	.23	.14
4	20	.47	.24	.39	.20
5	8	.63	.32	.42	.24
6	15	.47	.24	.23	.13
7	20	.59	.29	.49	.24
8	10	.61	.31	.41	.22
9	20	.45	.23	.38	.22
10	20	.44	.23	.48	.26
11	20	.49	.26	.24	.14
12	20	.43	.23	.32	.18
13	11	.37	.18	.31	.17

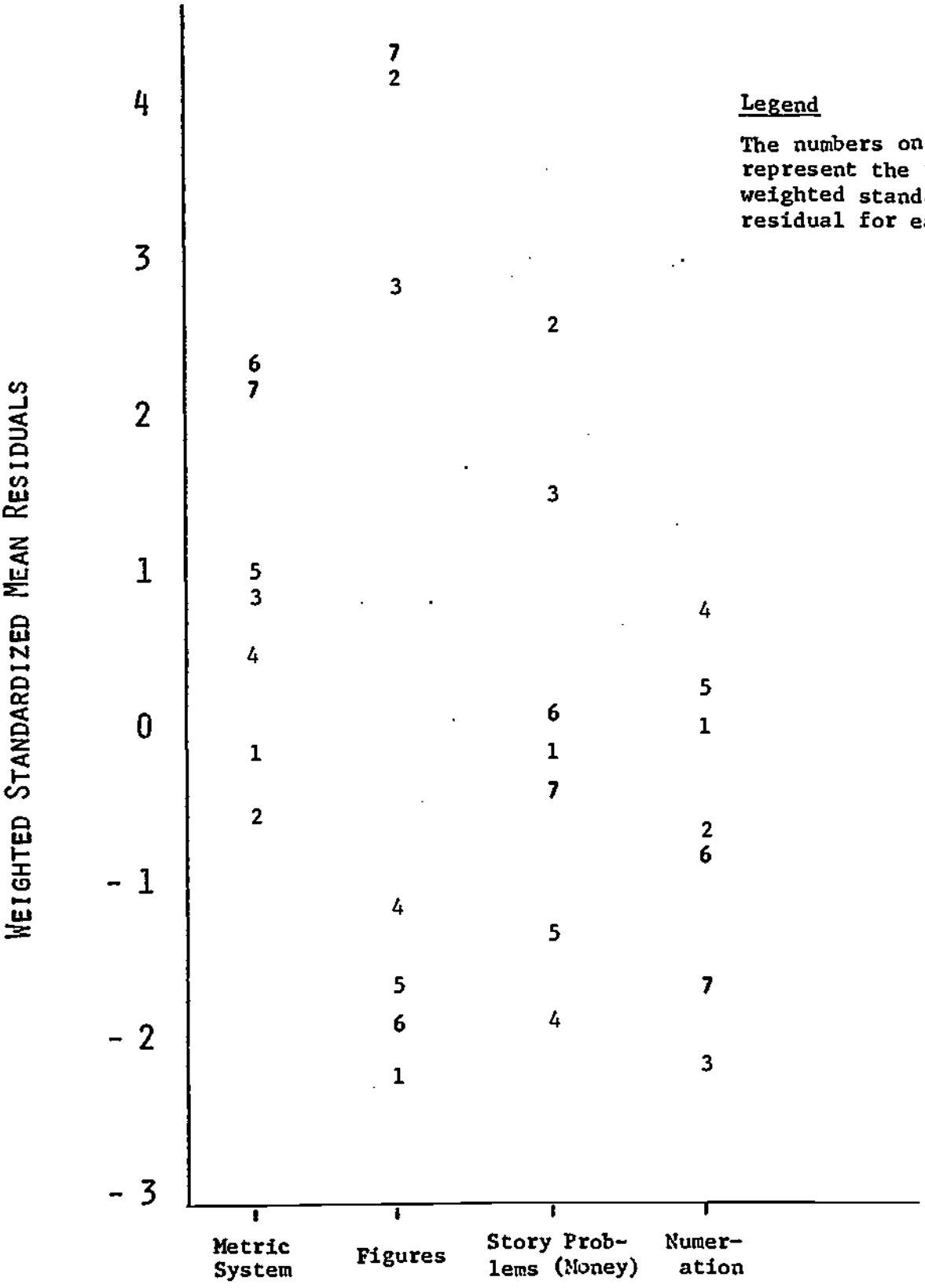
Table 7

Weighted Standardized Mean Residuals

Of Within School Item Difficulties by Content Category

	School						
Content Category	1	2	3	4	5	6	7
Calculation	-.07	-1.34	-.13	.11	.56	-.58	1.32
Definitions	1.27	.39	-.24	.22	.78	.00	-.32
Numeration	.00	-.66	-2.13	.78	.29	-.81	-1.64
Story Problems (general)	-.03	-.79	-.57	.63	-.76	.19	-.60
Story Problems (money)	-.10	2.60	1.51	-1.91	-1.29	.01	-.30
Metric System	-.10	-.56	.88	.56	1.08	2.28	2.20
Figures (fractions)	-2.29	4.16	2.94	-1.08	-1.64	-1.89	4.33
Unclassified	.93	-.85	.07	-1.44	1.61	.28	-.41

FIGURE 1



THE WEIGHTED STANDARDIZED MEAN RESIDUALS OF WITHIN SCHOOL ITEM DIFFICULTIES BY SELECTED CONTENT CATEGORY