DOCUMENT RESUME

ED 250 368 TM 840 679

AUTHOR	Allen, Thomas E.
TITLE	Out-of-Level Testing with the Stanford Achievement Test (Seventh Edition): A Procedure for Assigning Students to the Correct Battery Level.
INSTITUTION	Gallaudet Research Inst., Washington, DC.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE	Apr 84
GRANT	G0008300004
NOTE	45p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984). Small print in figure 9.
AVAILABLE FROM	Gallaudet Research Institute, 800 Florida Ave. NE Fay House, Washington, DC 20002 ($2.00, prepaid).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Achievement Tests; *Difficulty Level; Elementary Education; *Hearing Impairments; Item Analysis; Mathematics Achievement; Pilot Projects; Reading Achievement; Response Style (Tests); *Screening Tests; Test Construction; *Test Interpretation; Test Validity
IDENTIFIERS	*Out of Level Testing; *Stanford Achievement Tests; Stanford Achievement Tests for Hearing Impaired

ABSTRACT

In 1983, four screening tests for assigning students to the appropriate levels of the Stanford Achievement Test, Seventh Edition, were developed with a national sample of hearing impaired students. While students are normally assigned to one of six test level booklets according to grade, this is inappropriate for certain students. This paper describes: (1) the development of the screening tests; (2) the pilot testing and results; (3) the scoring system; and (4) the validity study of the screening tests using a norming sample of 8,331 hearing impaired students. Separate lower and upper level reading and mathematics tests, each containing approximately 30 items were constructed. The Sixth Edition of the Stanford was used as the criterion measure for assessing the discrimination power of the screening tests. The screening tests have elaborate scoring procedures, but result in excellent student placement into the appropriate levels of the Stanford Achievement Test, Seventh Edition. Response pattern analysis and individual item performance lead teachers to more in-depth consideration of test results. (BS)

Out-of-level testing with the Stanford Achievement Test (Seventh Edition): A procedure for assigning students to the correct battery level

Thomas E. Allen
Gallaudet Research Institute
Center for Assessment and Demographic Studies

Paper presented at the 1984 annual meeting of the American Educational Research Association, New Orleans, Louisiana

Out-of-level testing with the Stanford Achievement Test (Seventh Edition): a procedure for assigning students to the correct battery level

This paper will report on the development of a set of screening procedures for assigning students to the appropriate levels of the Stanford Achievement Test, Seventh Edition (Gardner, Rudman, Karlsen, & Merwin, 1982). Four screening tests were developed and piloted during the spring of 1982 with a national sample of hearing-impaired students, and the system for scoring the tests was developed after an analysis of the pilot data. The final tests were eventually used to screen over 8,000 students during the spring of 1983 when the Stanford was normed for the hearing-impaired student population. The screening tests form part of a set of special procedures and materials designed to facilitate the use of the Stanford with hearing-impaired students. This paper will describe the manner in which the screening tests were developed and piloted, present the results of the pilot testing, describe the scoring system that was developed, and report on the validity of the screening through a study of its use with the norming sample.

The Stanford Achievement Test is published in six difficulty levels (Primary 1, 2, and 3, Intermediate 1 and 2, and Advanced); each level is administered to hearing students in specific grades in school. The test booklets contain subtests in different content areas designed to test the progress of students with grade-appropriate material. Students are normally assigned to test booklets on the basis of their grade. The score information is then based on comparisons of the students' performance with the performance of students in the norming sample who were in the same grade when the tests were normed.

Relying on a student's grade or age as a basis for assignment to test level is often not appropriate. This is true for students whose progress in school lags significantly behind the progress of students who are similar in age or grade and for students whose growth in different achievement areas is uneven, i.e., they achieve at similar levels in some content areas, but lag behind in others. It is also inappropriate for students receiving instruction in programs with curricula which differ significantly from the curricula which guided the construction of the test.

Assigning a student to a level of the Stanford that is either too easy or too difficult leads to results that are not valid. For example, guessing on the Advanced level Reading Comprehension subtest can lead to a grade-equivalency estimate in the third to fourth grade range. Clearly, the value of this result is questionable. Norms such as these often become a part of students' permanent records, and, in the case of special education students, are used to make important planning decisions.

The need for quick and reliable procedures for determining appropriate test level assignments is great. Wick (1983) reported that, in 1974, 42% of the students in Chicago taking the Iowa Test of Basic Skills scored at the equivalent of a chance level, i.e., 25% or less in terms of their raw score. In some of the low-performance Chicago schools, the percentage was as high as 82%. This had the effect of elevating district averages when the raw scores were converted to norms. To solve the problem, Chicago switched to "functioning-level" test assignment, in which students were assigned to test level on the basis of "teacher opinion." Although this procedure led to a lower proportion of chance scores and a better test reliability, it is not clear what criteria teachers used in making their test level assignments. The project reported here was undertaken to develop a two-stage testing procedure in which a short screening test would provide the basis for making objective functional test level assignments.

In the current project, hearing-impaired students were used as the test development population. Assigning these students standardized achievement test levels on the basis of their age or grade is especially problematic. Allen, White, and Karchmer (1983) reviewed previous research findings related to the achievement levels of hearing-impaired students. They noted that the relationship between grade placement and skill level is often not the same for hearing-impaired students as it is for hearing students, and that hearing-impaired students' academic progress is uneven across content areas. They concluded that special procedures for assigning hearing-impaired students to levels of standardized tests are necessary. They also suggested that separate screenings in reading and math are necessary so that the subtests related to specific content areas are more adequately matched to the students' abilities. This population of students is one which has a need for special screening procedures if the results of standardized achievement testing are to be interpreted correctly.

METHOD

Test construction

Several guidelines were established to aid the construction of the screening tests:

1. Tests should be short, about 30 items each;

2. items selected for the screening tests should have a known statistical relationship in terms of their item difficulties to items that appear in the actual Stanford booklets;

3. separate screening tests in reading and math should be constructed;

4. items should be written in formats which are the same as formats used in the Stanford booklets;

5. lower and upper level screening tests should be constructed so that the range of ability levels measured by any one test would not be too wide;

6. the lower and upper levels should overlap in difficulty to allow for flexibility in assigning students to screening test levels who are achieving in the mid-range of ability.

The Psychological Corporation, publishers of the Stanford, made available to the current project the bank of test items which had been included in the initial item try-out for the Seventh Edition Stanford with a large national sample of hearing students. These items had been statistically analyzed along with the items that were selected for inclusion in the published edition of the test. Statistical information available for these items included biserial and point-biserial correlations, p-values for hearing students at different grade levels in the item try-out sample, and scale values of item difficulty, calculated through a Rasch analysis of the item data. Despite the fact that these items had been rejected from the set of items selected for the published test, there was an ample number of items available which had acceptable item statistics, i.e., biserial correlations above .40 and item difficulty indices which adequately represented the range of abilities measured by the different levels of the Stanford.

Means of the Rasch scale values of the items which had been selected by the publisher for publication in the Stanford were computed separately for the Mathematics Computation and Reading Comprehension subtests at each of the six levels. Where possible, items were selected for the screening tests from the remaining items which had scale values that clustered around these mean scores. This assured that the screening test items would adequately represent the entire range of ability measured across all six levels of the Stanford in the subject areas of reading comprehension and mathematics computation.

Each item in the bank was coded by the test authors to represent the

Stanford battery level for which it was being considered for inclusion. Using these codes to pick items for the screening instruments, eight items were selected to represent each of the six levels of Reading Comprehension, and eight items were selected to represent each of the six levels of Math Computation. The items were assembled into four booklets, each containing 32 items. Drafts of the booklets were sent to The Psychological Corporation for review and comment. The publisher noted some redundancy in the content of some of the items. As a result, several items were deleted from each of the booklets. An artist was employed to create the needed artwork for the booklet in a style that was consistent with that used by the test publisher in creating the final forms of the test. The final versions of the screening tests were constructed as follows:

Form R1A -Lower Level Screening Test in Reading, containing items from Primary 1, Primary 2, Primary 3, and Intermediate 1 Reading Comprehension subtests (27 items).

Form R2A -Upper Level Screening Test in Reading, containing items from Primary 3, Intermediate 1, Intermediate 2, and Advanced Comprehension subtests (30 items).

Form M1A -Lower Level Screening Test in Mathematics, containing items from Primary 1, Primary 2, Primary 3, and Intermediate 1 Mathematics Computation subtests (26 items).

Form M2A -Upper Level Screening Test in Mathematics, containing items from Primary 3, Intermediate 1, Intermediate 2, and Advanced

Mathematics Computation subtests (26 items).

Samples

Development sample. Students selected for inclusion in the pilot testing project were drawn from the population of students on whom data had been collected by the Annual Survey of Hearing Impaired Children and Youth (AS) during the spring of 1981. This survey collects information yearly on over 55,000 hearing-impaired students who receive special education services in programs throughout the United States. Nearly 1,100 programs containing over 5,000 individual schools throughout the country participate in this survey every year.

A random sample of schools was selected from the AS data base to represent the different regions of the country and the different types of educational programs serving hearing-impaired students. A total of 84 schools throughout the country participated in the project. Of these, 76 schools completed all the required testing. The total number of students tested in these schools was 1,450.

Verification sample. The screening procedures developed during the first year of this project were used in the following year to assign hearing-impaired students to the six levels of the Stanford when the test was normed on a large national sample of hearing-impaired students. The screening tests were administered to 8,331 hearing-impaired students, chosen through a random sampling of the programs which participate in the Annual Survey.

Design

Criterion measure. During the year in which the pilot testing was being carried out, the 7th Edition of the Stanford was not available in its final form. The 6th Edition of the Stanford was therefore used as the criterion measure for assessing the discriminating power of the new screening tests. This procedure was considered satisfactory since the grade-level to battery-level relationship is approximately the same for both the 6th and 7th editions of the Stanford.

During the 1973-74 school year the 6th Edition of the Stanford Achievement Test was normed on a large national sample of hearing-impaired students (Office of Demographic Studies, 1974). During that project, the problems of functional-level versus grade-level test assignment were also addressed. The result was a modified version of the Stanford called the Special Edition for Hearing-Impaired Students of the Stanford Achievement Test (SAT-HI). It is important to consider two features of this special edition in the present design:

1. The Reading Comprehension subtests from the Form B Primary 2 and Intermediate 1 levels of the Stanford served as upper and lower level screening tests for the Form A batteries. There was no separate screening for math.

2. To get around the uneven growth problem, the test booklets were reconstructed, i.e., subtests from different Stanford battery levels were mixed, and special booklets were printed to approximate

the median growth patterns of hearing-impaired students in the different subtest areas. Six levels of the SAT-HI were constructed. The Reading Comprehension and Mathematics Computation subtests included in each of these levels are as follows:

SAT-HI Level 1	P1 Reading	P1 - Mathematics
SAT-HI Level 2	P2 Reading	P3 - Mathematics
SAT-HI Level 3	P3 Reading	I1 - Mathematics
SAT-HI Level 4	I1 Reading	I2 - Mathematics
SAT-HI Level 5	I2 Reading	Ad - Mathematics
SAT-HI Level 6	Ad Reading	Ad - Mathematics

The problem posed by using the SAT-HI as the criterion measure was that the Primary 2 Mathematics Computation subtest is never administered. In determining cut-off scores for assignment to the Primary 2 Mathematics Computation level, a pseudo Primary 2 math criterion group was created through interpolation. This procedure is discussed below.

Test assignments and criterion groups. Students in the pilot project were first administered the screening tests designed for use with the 6th Edition. These were hand-scored by the teachers participating in the project, and, as a result, students were assigned to one of six levels of the SAT-HI. These level assignments defined six criterion groups for studying the new screening instruments. In the analysis, these groups will be referred to as criterion groups 1 through 6, rather than Primary 1 through Advanced, since the SAT-HI combines subtests from various Stanford levels within each of its own levels.

Soon after the SAT-HI level assignments were made, students were assigned separately to different levels of the new reading and math screening tests. Teachers were asked to make independent judgments as to whether they felt each student was above or below the fifth grade level in reading and math. For hearing students the fifth grade level is roughly the dividing point for assignment to the Primary 3 and Intermediate 1 test booklet levels. Hearing-impaired students in the current sample who were judged to be at or above the fifth grade level in either reading or math were administered the appropriate upper level screening test (Form R2A or Form M2A). Students judged to be below the fifth grade level in either reading or math were administered the appropriate lower level screening test (Form M1A or R1A).

Each student took a total of four tests: the screening test used with the 6th Edition SAT-HI; one of six levels of the SAT-HI; either Form R1A or R2A (determined by the teacher's opinion of the student's reading ability); and either Form M1A or M2A (determined by the teacher's opinion of the student's mathematics ability).

Validation. When the Stanford was normed on a national sample of hearing-impaired students in the spring of 1983, the screening procedures developed the previous year were used to assign students to test levels.

To assure that the screening procedures were rigorously followed, all screening tests were computer scored by the norming project office.

For the Reading Comprehension and Mathematics Computation subtests at each of the six levels, acceptable raw score ranges were determined:- 25% of the total number of items as the lower boundary and 90% of the total number of items as the upper boundary. Students scoring within this range were judged as having interpretable or acceptable scores. (Only students whose actual test level matched the assigned level were studied in this part of the analysis. Approximately 5% of the norming sample were either not screened or were administered a level of the test which differed from the level suggested by the screening test results).

RESULTS

Table 1 shows the means and standard deviations of raw scores on the four screening tests for each of the six criterion groups defined by the SAT-HI test level assignments. It also shows estimates of the test reliabilities, computed using the KR-20 formula. Students who screened into levels 1 and 2 of the SAT-HI using the 1974 screening procedures, but who were rated as being <u>above</u> the fifth grade level by their teachers (and were therefore assigned to the upper level screening tests), were excluded from this analysis. Also excluded were students who screened into levels 5 and 6 of the SAT-HI, but who were judged to be <u>below</u> the fifth grade level by their teachers (and were therefore assigned to the lower level screening tests). These students were excluded since they took levels of the SAT-HI which were not represented by items included in the screening tests to which they were assigned. When these students were excluded, the resulting sample consisted of 1,374 students who took both the SAT-HI Reading Comprehension subtest and a reading screening test, and 1,357 students who took both the SAT-HI Mathematics Computation subtest and a math screening test.

Insert Table 1 here.

The means in Table 1 give some idea of the discriminating power of the new tests. The mean raw scores on Form R1A for criterion groups 1, 2, and 3 are markedly different, with jumps of over 4 points at each successive level. Criterion groups 3 and 4 differed in their mean performance on Form R1A by only 1.4 points. While the students in criterion group 4 were assigned by the old screening procedures to take the Intermediate 1 Reading Comprehension test, their teachers rated their ability below the fifth grade level. Thus we should not expect their performance on Form R1A to differ dramatically from the performance shown by group 3.

Form R2A does less well discriminating the upper level criterion groups, as can be noted by the mean values for Form R2A in Table 1. The difference between means for groups 4 and 5 is particularly small (2.2

points).

Form M1A shows a pattern for criterion groups 1-4 in math similar to the pattern noted for this same group in reading. Criterion groups 1, 2 and 3 were well differentiated, while groups 3 and 4 had almost identical mean scores. Criterion groups 1 and 2 differed in mean raw score performance by a large 6.1 points. (Students who took level 2 of the SAT-HI actually took the Primary 3 Math Computation subtest.) The large difference in screening test performance by criterion groups 1 and 2 shows that hearing-impaired students progress in math at a faster rate than they do in reading. These results confirm the necessity for separate screenings in math and reading. Form M2A shows the least discriminating power of all the four tests. Criterion groups 4, 5, and 6 had mean raw scores that were all very close. Since groups 5 and 6 were both assigned to the Advanced level of the Mathematics Computation subtest, we would not expect these two groups to differ markedly on their screening test performance.

The reliabilities were all over .80. The two lower level tests which had higher variability (and better discriminability among the criterion groups) showed slightly higher reliability than the two upper level tests which were more restricted in range.

Insert Figures 1-4 here

Figures 1 to 4 show the discriminating features of the four screening tests more clearly. In these figures, the cumulative relative distribution are plotted for all criterion groups for each of the four screening tests. For these plots, the criterion groups were restricted to students scoring in the inter-quartile range of the appropriate SAT-HI subtests. These students are the ones who are the most ideally placed in terms of Stanford test level assignment.

Figures 1 to 4 confirm the mean score findings: Forms R1A and M1A were good discriminators of students taking levels 1, 2 and 3 of the SAT-HI. Level 4 performance on Form R1A was not distinguishable from level 3 performance. (The criterion group 4 performance on Form M1A is not plotted since the inter-quartile range for this group only contained 21 students. Also, criterion group 4 took the Intermediate 2 math subtest, which is not represented by the Form M1A screening test items.)

The upper level screening tests had less discriminating power. In reading, the distinction between criterion groups 4 and 5 (Intermediate 1 and Intermediate 2 assignments, respectively) was very slight. In math, the distinction between criterion groups 3 and 4 (also Intermediate 1 and Intermediate 2 assignments) was equally poor.

Scoring

The goal of the scoring system that was developed was to give teachers a way to assign students to levels of the Stanford test battery in reading and math. The results of the reading screening test should help teachers assign their students to the reading and reading-related subtests in the

Stanford battery. The results of the math screening test should help teachers assign their students to the appropriate levels of the math subtests.

The analysis above revealed that students taking different levels of the Stanford, especially those taking the lower three levels, performed differently on the screening tests. Nonetheless, the following facts also had to be taken into account:

1) Although the distributions of screening test scores differed for the different criterion groups, there was considerable overlap, especially at the upper levels.

2) Because the Stanford may not be ideally suited for all hearing-impaired students, and because the screening tests were so short, some study of the response patterns of the test takers was necessary to assure teachers of the validity of the assignments. A procedure was needed which allowed teachers to study the individual response patterns.

Score ranges and border regions. The screening test raw score ranges for students who scored in the middle 50% of each criterion group were determined. These ranges are plotted in Figures 5 through 8 for the four screening tests. Border regions were defined as the raw score values which were included in the mid-ranges of two different criterion groups. These border regions are also indicated on Figures 5 through 8.

Insert Figures 5-8 here

In Figures 5 through 8 the actual Stanford test levels are indicated for each criterion group. Figure 7 shows the interpolated Primary 2 criterion group for Form M1A. This interpolation was necessitated by the subtest structure of the SAT-HI, in which the Primary 2 Math Computation subtest is not administered. The Primary 1 and Primary 3 criterion groups overlapped only at the raw score value of 15. A pseudo-Primary 2 criterion group was created which was defined by 15 plus and minus 2. This interpolation resulted in a Primary 1 to Primary 2 border region and a Primary 2 to Primary 3 border region, as shown in Figure 7.

Scoring rules related to border regions. When students do not score in a border region, their test level assignment is determined by the criterion group range in which they fall. Students who score in a border region could be assigned to either of the adjacent test levels. To help teachers decide which of the two adjacent levels is the most appropriate, a table of "Best Discriminating Items" was developed.

Insert Tables 2-5 here

The "Best Discriminating Items" are those items which are the best discriminators between two adjacent test levels. To determine which items were the best discriminators, p-values were computed for each item for each criterion group. Then, p-value differences were computed for adjacent levels. These p-value differences are shown in Tables 2 through 5 for the four screening tests. The 7.6 shown as the Primary 1 to Primary 2 p-value difference for Form R1A indicates that 7.6% more of the students in the Primary 2 criterion group answered item 1 correctly than answered it correctly in the Primary 1 criterion group.

For each of the adjacent levels, the four best discriminating items were noted. These were the items that had the largest p-value differences for the adjacent levels.

When students score in a border region, teachers are asked to look more carefully at the best discriminating items. If students have answered at least three of the four best discriminating items correctly, they should be assigned to the higher of the two adjacent levels. If they fail to answer at least three of the four best discriminating items correctly, they should be assigned to the lower of the two adjacent levels.

Response pattern assessment. The items selected for the screening tests have a known statistical relationship to the items published in the Stanford battery. The Rasch scaled difficulty values of these items place them in the context of the reading comprehension and mathematics computation scales that have been developed for the six-level battery. An important component of the screening process is to identify students who respond to these items in a way that violates the assumptions of the scale, i.e., that the items are hierarchically arranged along a unidimensional scale.

For special populations such as hearing-impaired students, a check on how well the scale "fits" the students is crucial. If special education students attend special programs, it is possible that their curricula are not well represented by the test items. Also, they may show special growth patterns in which the hierarchy of skills is acquired in a different sequence. Finally, with short tests, guessing poses a problem unless the pattern of item responses is taken into consideration.

Much of the score information from the Stanford is based on raw score conversions. The legitimacy of these conversions depends on a good fit between the student and the scale. The current scoring procedures sought to provide information to teachers about the response patterns of their students who showed unusual patterns of item responses.

Special scoring sheets were developed to enable teachers to study the response patterns of their students. (See Figure 9.) On these sheets, grids were printed which rearranged the items by the Rasch item-difficulty indices provided by the test publishers. Teachers are instructed on these sheets to transfer the student responses to the grid. This enables them to study each student's pattern of item responses. Ideally, each student should answer correctly all items which have a difficulty ranking equal to and less than their raw score. More care should be given in assigning students who answer a substantial number of items correctly which have difficulty rankings above their raw score. These students may have guessed well, or they may not be well suited for testing by the Stanford.

Criterion for identifying unusual response patterns. Standard errors for each of the four screening tests were within two raw score points. Therefore, the procedures instruct teachers to consider correct item responses unusual only if their difficulty ranking is greater than 4 positions (two standard errors) above the obtained raw score. Teachers then count up the number of unusual responses and divide that number by the raw score. If the total number of items correct (the raw score) is comprised of more than 30% unusual correct responses, then the student should receive special consideration before the test level is assigned.

Scoring rules for students with unusual response patterns. Students whose raw score is comprised of a large number (> 30%) of unusual correct responses are difficult to assign to appropriate levels of the Stanford. There are several reasons why they may have responded in an unusual fashion to the screening test. They may have guessed well; their curriculum may not match the test; their growth patterns may be such that they develop skills in a different sequence. The following rule was devised as a practical solution to the problem of assigning these students: Reduce their raw scores to the next lowest border region and apply the best discriminating items test to their responses. While this procedure does not guarantee that students will be correctly assigned, it forces teachers to consider a subset of items which have good discriminating power between different test levels.

Summary of the scoring procedure. To score the new screening tests, the following procedure is used:

1. Transfer item responses to the scoring sheet.
2. Score the items. Calculate the raw score.
3. Determine if raw score is comprised of more than 30% "unusual" correct responses.
4. Determine if raw score is in a border region.
5. If step 3 is true, reduce raw score to the next lowest border region.
6. If step 4 is true, or if the raw score has been reduced because of an unusual response pattern, apply the appropriate discriminating item test to assign test level.
7. If neither step 3 nor 4 is true, use the obtained raw score to assign test level.

The scoring sheets which contain the rearranged item grids also contain instructions for completing all of the steps listed above. The sheet developed for Form R1A appears in Figure 9.

Insert Figure 9 here

Administering a single screening test to each student will result in each student being placed into one of nine categories with a separate assignment or special instruction for each, as follows:

1. Scored too low on the lower level screening test. Achievement level is perhaps too low for entry level into the battery.
2. Assign to Primary 1.
3. Assign to Primary 2.
4. Assign to Primary 3.
5. Scored too high on the lower level screening test. Administer upper level test before making assignment.
6. Scored too low on upper level screening test. Administer lower level test before making assignment.
7. Assign to Intermediate 1.
8. Assign to Intermediate 2.
9. Assign to Advanced

Validation of screening procedures

Insert Table 6 here

Table 6 shows the proportions of students from the norming sample who scored in each of three different raw score ranges at each level of the Stanford. These ranges are 1) <26% of the items correct (chance level); 2) 26% to 90% of the items correct (acceptable level); and 3) >90% of the items correct (top-out level).

All of these students were assigned to their test levels using the procedures described above. The total number of students in this table does not equal the 8,331 tested in the norming because only the students who were classified into categories 2, 3, 4, 7, 8 and 9 are reported. Due to time constraints, students in the norming sample who scored too high on the lower level screeners or too low on the upper level screeners could not be re-screened. They were assigned to the next highest or lowest levels, respectively, but are not reported in Table 6. Students who scored too low on the lower level screeners (category 1) were assigned to Primary 1. These students are also not included in Table 6.

For Reading Comprehension 96% of the sample scored in an acceptable range. This percentage is fairly consistent across all levels of the test. There is a slightly higher likelihood for students assigned to Primary 1 to score in the top-out category (3.1% compared with 1.0% overall), and for students at the Intermediate 2 and Advanced levels to score at chance level (4.5% and 5.0% compared with 2.1% overall). However, these percentages are quite small. The screening tests placed an overwhelming majority of students into a correct reading level.

For Math Computation, 83.6% of the sample scored in an acceptable range. Only 1.0% of the students scored at chance level, and 15.4% score in the top-out category. These results imply that the computational abilities of 15% of the students in the norming sample were underestimated by the screening tests.

In the math area, it is useful to consider other subtests which are assigned on the basis of the math screening test. The special procedures developed for using the Stanford with hearing-impaired students recommend assigning the Math Applications subtest on the basis of the reading screening since the test requires considerable verbal ability, and hearing-impaired students tend to perform at a lower verbal level than math level. The Concepts of Number subtest, on the other hand, is assigned on the basis of the math screening test. It is useful to consider the Concepts of Number raw scores obtained by the norming sample at each level of the battery..

Insert Table 7 here

Table 7 shows the proportions of students who scored in each of the three performance categories for Concepts of Number. These data show that, for Concepts of Number, 94.2% of the sample scored in a acceptable range. Approximately 3% scored at chance level and 2.5% scored in the top-out level. Thus, while a fairly high proportion of students top-out of the Math Computation subtest, the proportion is much lower for Concepts of Number. Since students take both subtests in the level determined by the math screening test, these results are encouraging.

CONCLUSION

The screening tests developed in this project have elaborate scoring procedures. Nonetheless, when followed carefully, they result in excellent placements of students into appropriate levels of the 7th Edition of the Stanford Achievement Test.

A side-effect of the scoring procedure is that it leads teachers to consider test results in a more in-depth manner than simply converting a raw score to a test level assignment. They are encouraged to consider the response patterns of individual students as valuable sources of information. They are led to consider situations where students score in border regions. They are forced to look at performance on individual items as input to important decisions. It is hoped that the teachers who use these procedures will develop sophistication and that they will therefore approach test results with a more critical eye. Response pattern analysis and consideration of individual item performance are not activities that are reserved for screening tests alone.

REFERENCES

Allen, T.E., White, C.S., & Karchmer, M.A. (1983) Issues in the development of a special edition for hearing-impaired students of the Seventh Edition of the Stanford Achievement Test. American Annals of the Deaf, 128, 34-39.

Gardner, E.F., Rudman, H.C., Kaelsen, B., & Merwin, J. (1982) Stanford Achievement Test, 7th Edition. New York: Harcourt Brace Jovanovich.

Office of Demographic Studies. (1974) Score conversion tables and age-based percentile norms for Stanford Achievement Test, Special Edition for Hearing Impaired Students. Washington, DC: Gallaudet College.

Wick, J.W. (1983) Reducing proportion of chance scores in inner-city standardized testing results: Impact on average scores. American Educational Research Journal, 20, 461-463.

Author Notes

Preparation of this paper was supported in part by the Gallaudet Research Institute and by grant # G0008300004, U.S. Office of Education, Special Education Programs.

The author gratefully acknowledges the assistance of Tamara Osborn and Alex Quaynor in the collection and analysis of data and of Michael Karchmer and Arthur Schildroth for comments on a draft of this article.

Requests for reprints should be sent to the author, Center for Assessment and Demographic Studies, Gallaudet Research Institute, 800 Florida Ave. NE, Washington, D.C.

Table 1

SCREENING TEST MEANS AND STANDARD DEVIATIONS
BROKEN DOWN BY SAT-HI TEST LEVEL POPULATIONS

SAT-HI		FORM	FORM	FORM	FORM
LEVEL		R1A	R2A	M1A	M2A
1	X	10.2	----	11.8	----
	SD	3.79		5.54	
	N	274		272	
2	X	14.9	----	17.9	----
	SD	4.66		5.14	
	N	335		294	
3	X	19.5	16.5	20.9	16.9
	SD	4.12	4.56	4.13	3.77
	N	266	53	162	169
4	X	20.9	19.7	21.0	19.4
	SD	4.54	4.08	3.85	3.31
	N	90	100	42	148
5	X	----	21.9	----	20.3
	SD		3.79		3.38
	N		121		132
6	X	----	25.7	----	21.3
	SD		2.89		3.08
	N		135		138
TOTAL Ns		965	409	770	587
RELIABILITY: KR-20		.87	.83	.92	.81

Table 2

FORM R1A
P VALUE DIFFERENCES IN ADJACENT
TEST LEVELS

ITEMS	P1 TO P2	P2 TO P3	P3 TO I1
1	7.6	1.5	1.9
2	3.3	1.0	-1.4
3	5.4	4.2	8.6
4	26.4	8.1	-1.3
5	23.6	7.7	3.0
6	37.2*	6.7	-2.6
7	33.2*	26.1	-4.6
8	33.0*	11.5	-0.2
9	21.6	15.9	5.4
10	37.3*	12.4	1.6
11	16.5	26.6	8.9
12	28.7	24.6	4.8
13	28.5	6.5	9.5
14	-5.5	15.8	20.0*
15	28.6	23.8	1.0
16	12.7	11.7	17.7*
17	5.3	18.3	-3.5
18	-1.3	-0.7	19.4*
19	20.8	23.4	6.3
20	18.1	22.9	-8.7
21	14.7	30.6*	7.6
22	18.8	17.0	4.8
23	5.2	30.4*	6.6
24	16.6	29.4*	0.6
25	23.2	30.5*	4.8
26	7.8	24.0	17.8*
27	14.5	16.5	11.5

*Best Discriminating Items

Table 3

FORM R2A
P VALUE DIFFERENCES IN ADJACENT
TEST LEVELS

ITEMS	P3 TO I1	I1 TO I2	I2 TO ADV
1	8.6	-3.1	3.4
2	22.2*	14.6	4.4
3	11.8	7.4	8.6
4	18.8	11.9	14.2
5	13.8	8.4	5.9
6	2.3	6.7	4.6
7	17.1	0.2	1.4
8	6.8	8.8	8.0
9	11.2	3.4	2.2
10	8.2	-0.2	2.8
11	13.4	1.2	-0.7
12	13.8	8.4	4.4
13	24.3*	14.9*	22.2
14	-11.0	10.4	6.3
15	7.8	10.8	18.7
16	-3.7	16.5*	25.6*
17	20.5*	5.5	22.2
18	7.6	6.5	24.0
19	-6.5	17.1*	25.5*
20	6.1	8.1	5.8
21	4.8	3.7	10.4
22	-7.2	15.5*	21.8
23	-6.8	-1.5	18.9
24	1.6	5.7	13.2
25	11.8	10.9	16.8
26	16.1	-8.3	38.6*
27	10.1	9.1	5.9
28	19.0	11.4	12.7
29	19.8*	9.6	10.1
30	18.2	5.5	26.2*

*Best Discriminating Items

Table 4

FORM M1A
P VALUE DIFFERENCES IN ADJACENT
TEST LEVELS

ITEMS	P1 TO P3	P3 TO I1	I1 TO I2
1	14.4	0	3.7
2	17.5	1.6	-2.2
3	13.6	2.2	0.8
4	22.1	-0.4	-0.9
5	28.5	3.8	-3.8
6	24.0	3.2	3.2
7	18.0	8.5	6.9
8	21.5	5.4	10.6*
9	12.7	10.2	6.5
10	34.0	6.8	-3.8
11	25.4	15.4	-3.5
12	24.1	9.3	1.2
13	35.5*	13.8	-10.4
14	38.6*	9.2	4.4
15	23.7	6.7	-4.0
16	33.6	14.6	-9.9
17	37.1*	21.1	8.9*
18	27.6	15.3	-7.1
19	30.3	10.9	0.3
20	35.6*	18.7	-1.0
21	23.3	25.0*	-6.0
22	21.0	14.6	-6.1
23	-4.8	20.0	-3.3
24	8.9	24.7*	7.6*
25	15.7	22.7*	2.8
26	21.0	22.1*	9.5*

*Best Discriminating Items

Table 5

FORM M2A
P VALUE DIFFERENCES IN ADJACENT
TEST LEVELS

ITEMS	I1 TO I2	I2 TO ADV
1	3.0	-2.0
2	-0.9	-1.0
3	-2.6	4.1
4	2.8	-1.0
5	10.9	1.8
6	1.8	2.6
7	6.8	1.4
8	14.6	10.2
9	1.6	2.1
10	19.5*	2.6
11	13.4	5.2
12	11.5	-1.0
13	7.2	12.2
14	10.2	-2.1
15	9.7	-0.2
16	5.1	-0.6
17	7.7	-1.3
18	17.7*	10.8
19	12.2	5.7
20	7.7	19.2*
21	9.5	16.9*
22	20.3*	4.6
23	12.3	17.3*
24	27.0*	15.6*
25	11.7	14.1
26	6.4	1.9

*Best Discriminating Items

Table 6

PERCENT SCORING IN EACH OF THREE PERFORMANCE CATEGORIES FOR READING COMPREHENSION AND MATH COMPUTATION AT EACH OF THE SIX STANFORD ACHIEVEMENT TEST BATTERY LEVELS

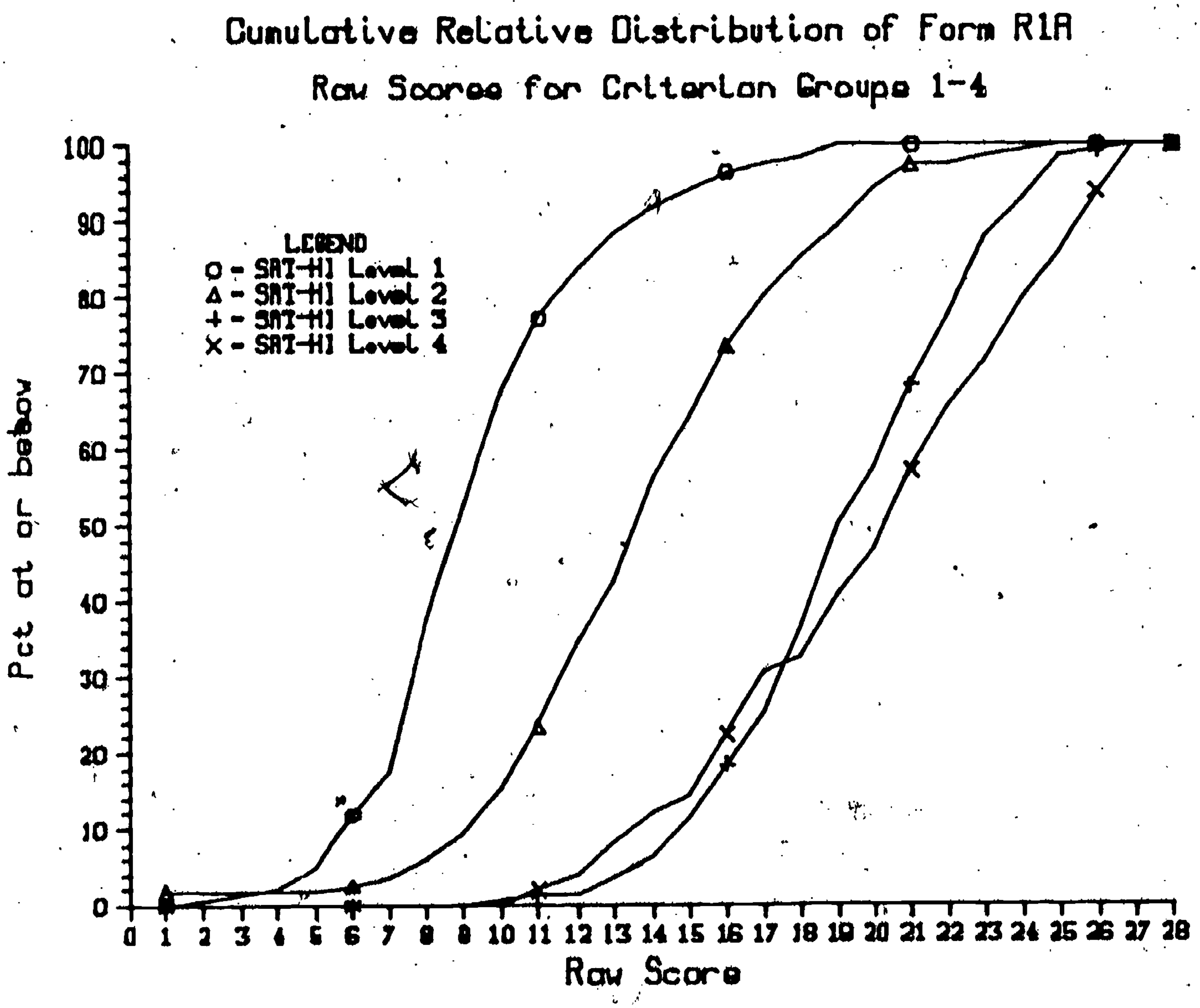
	N	Chance <26%	Acceptable 26% - 90%	Top-out >90%
Reading Comprehension				
Primary 1	1335	0.9%	96.0%	3.1%
Primary 2	1694	2.3%	97.6%	0.1%
Primary 3	1788	1.3%	98.6%	0.1%
Interm. 1	455	1.3%	98.5%	0.1%
Interm. 2	268	4.5%	95.1%	0.4%
Advanced	959	5.0%	93.7%	1.3%
Overall	6499	2.1%	96.9%	1.0%
Mathematics Computation				
Primary 1	958	1.6%	76.1%	22.3%
Primary 2	516	0.0%	88.0%	12.0%
Primary 3	1399	1.1%	77.3%	21.6%
Interm. 1	1648	1.1%	85.9%	13.0%
Interm. 2	1094	0.5%	83.9%	15.6%
Advanced	1178	0.9%	91.7%	7.4%
Overall	6793	1.0%	83.6%	15.4%

Table 7

PERCENT SCORING IN EACH OF THREE PERFORMANCE CATEGORIES
FOR CONCEPTS OF NUMBER AT EACH OF THE SIX
STANFORD ACHIEVEMENT TEST BATTERY LEVELS

	N	Chance <26%	Acceptable 26% - 90%	Top-Out >90%
CONCEPTS OF NUMBER				
Primary 1	954	0.8%	95.8%	3.4%
Primary 2	522	1.0%	96.3%	2.7%
Primary 3	1398	2.5%	96.2%	1.3%
Interm. 1	1653	8.2%	90.2%	1.6%
Interm. 2	1091	2.0%	96.8%	1.2%
Advanced	1177	1.5%	93.1%	5.4%
Overall	6795	3.3%	94.2%	2.5%

Figure 1


Cumulative Relative Distribution of Form R1A
Raw Scores for Criterion Groups 1-4
LEGEND
O - SAT-HI Level 1
Δ - SAT-HI Level 2
+ - SAT-HI Level 3
X - SAT-HI Level 4
Pct at or below
0
10
20
30
40
50
60
70
80
90
100
Raw Score
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28

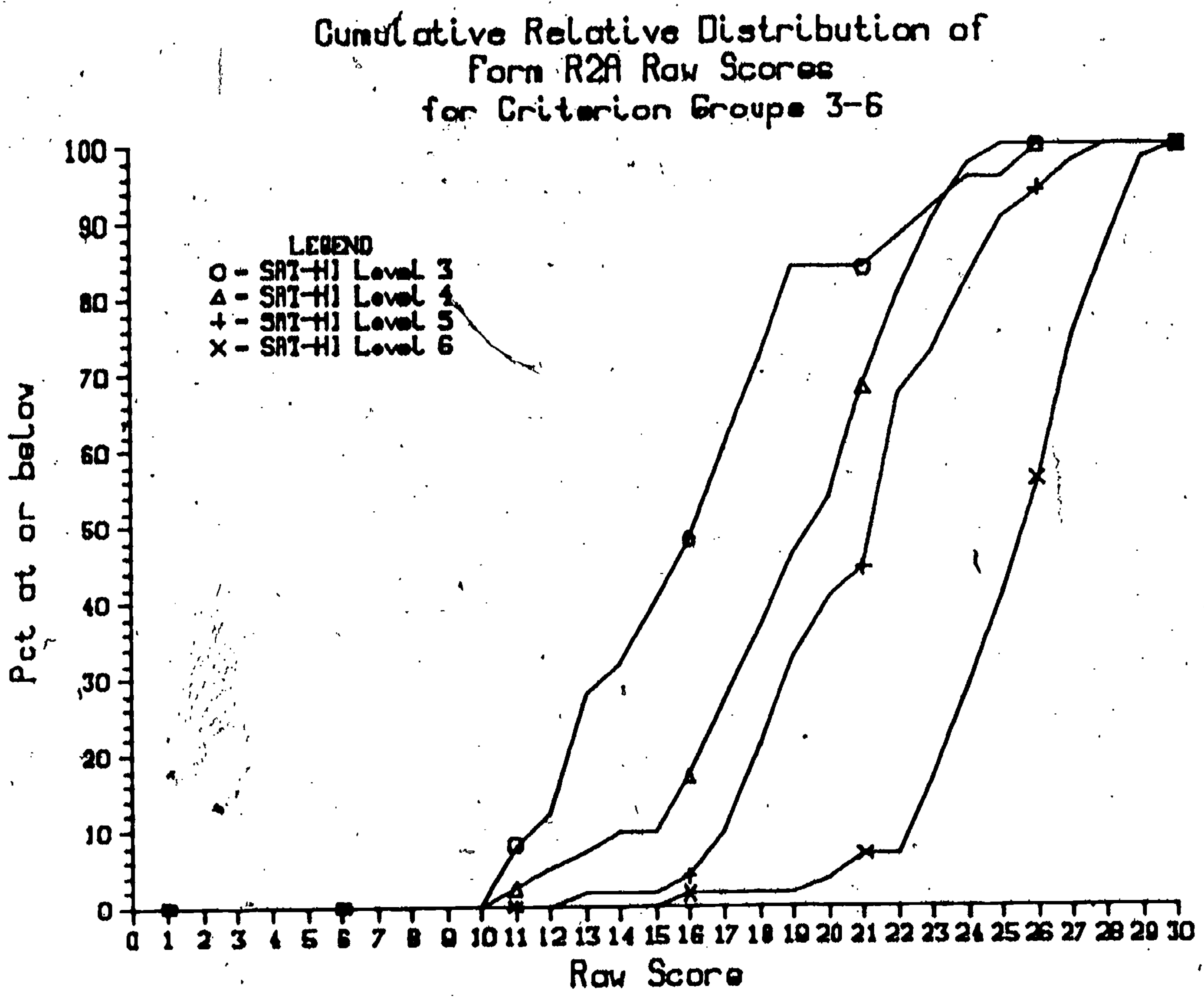

Cumulative Relative Distribution of
Form R2A Raw Scores
for Criterion Groups 3-6
LEGEND
○ - SAT-HI Level 3
△ - SAT-HI Level 4
+ - SAT-HI Level 5
× - SAT-HI Level 6
Pct at or below
0
10
20
30
40
50
60
70
80
90
100
Raw Score
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Figure 3

Cumulative Relative Distribution of Form M1A
Raw Scores for Criterion Groups 1-3

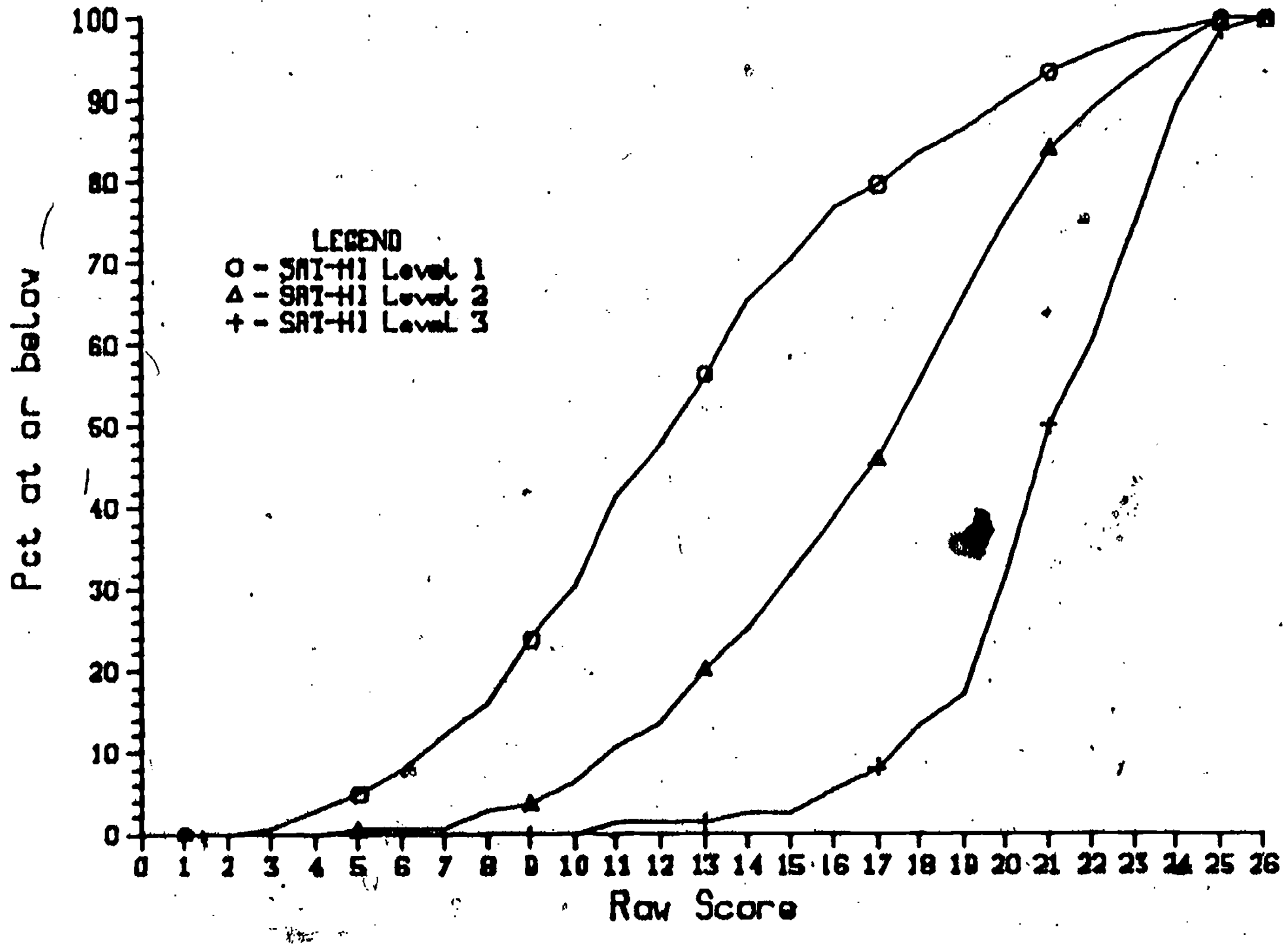


Figure 4

Comulative Relative Distribution of Form M2A
Raw Scores for Criterion Groups 3-6

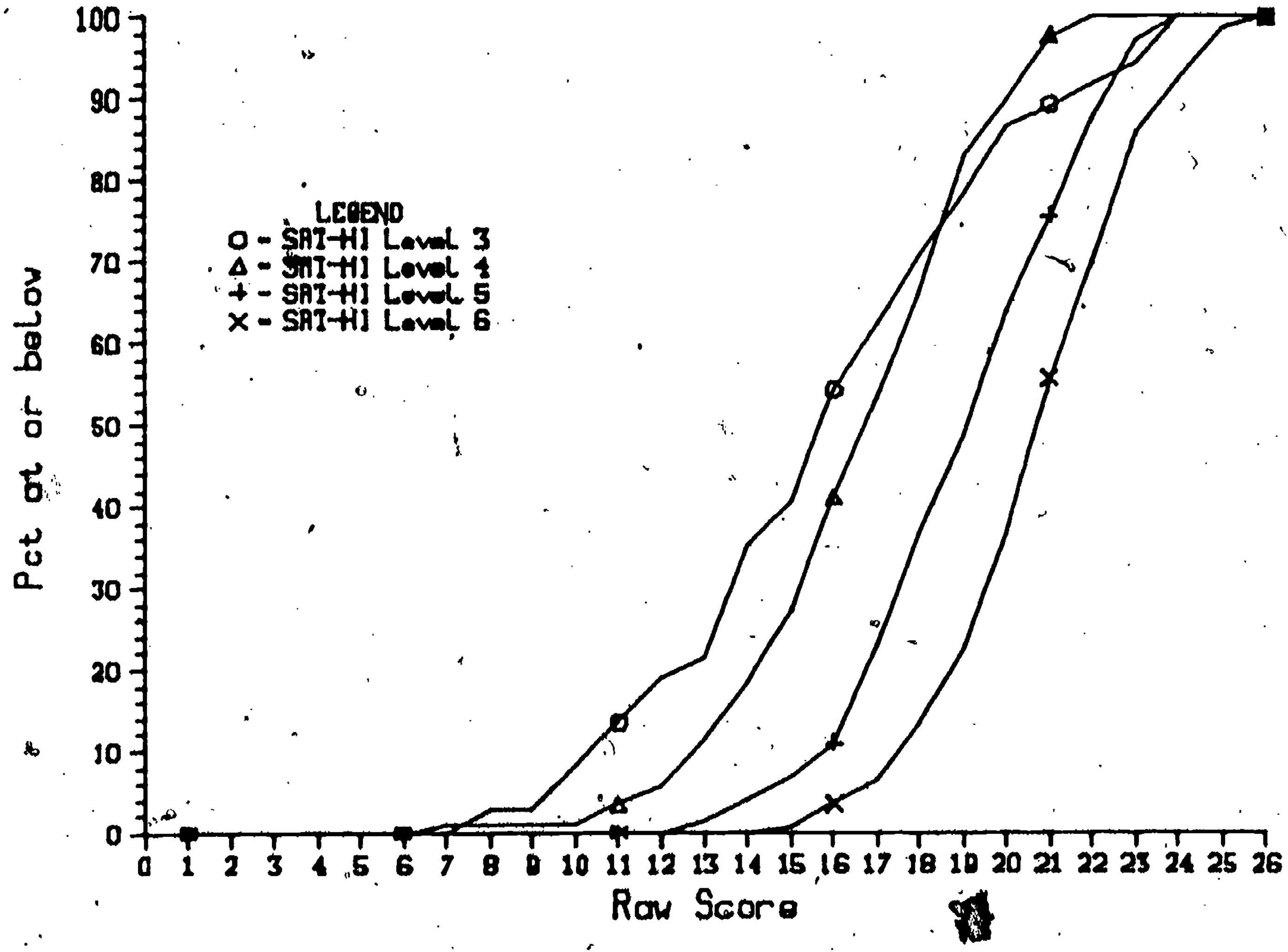


Figure 5

Interquartile Ranges of Form R1A
Raw Scores for Primary 1, Primary 2, and
Primary 3 Criterion Groups

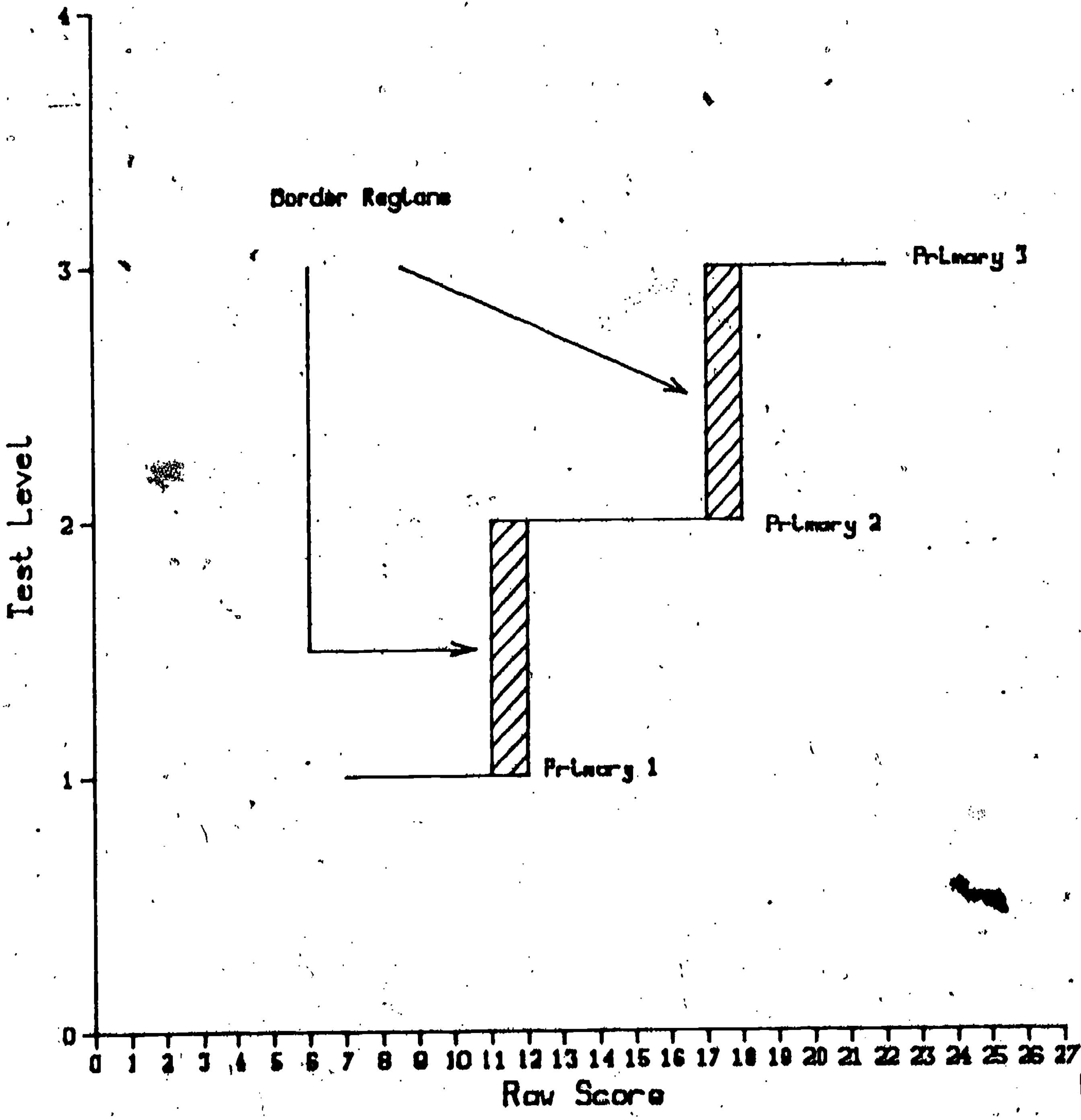


Figure 6

Interquartile Ranges of Form R2A Raw Scores for Primary 3, Intermediate 1, Intermediate 2, and Advanced Criterion Groups

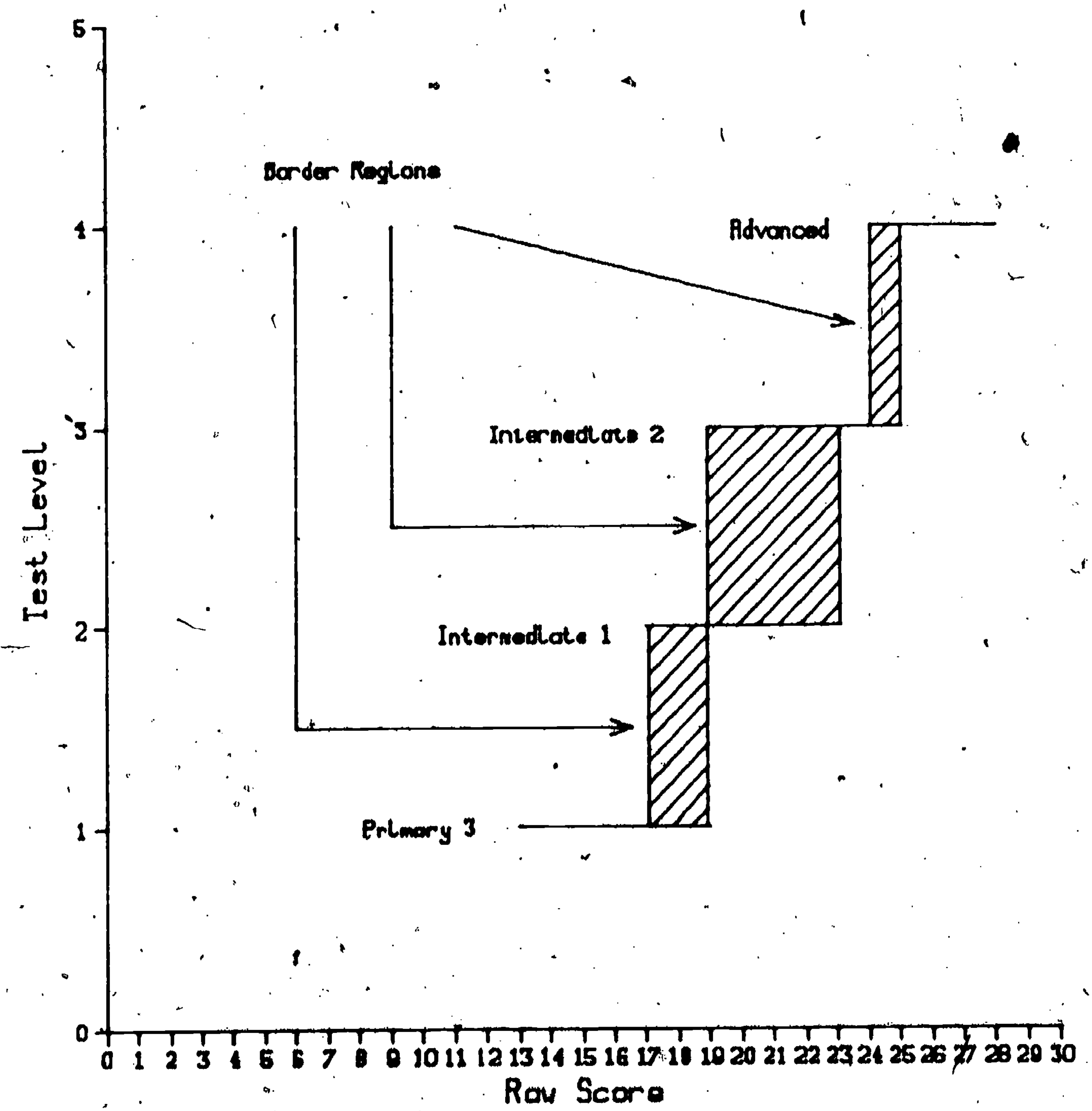


Figure 7

Interquartile Ranges of Form M1A Raw Scores
for Primary 1, Primary 2, Primary 3, and
Intermediate 1 Criterion Groups

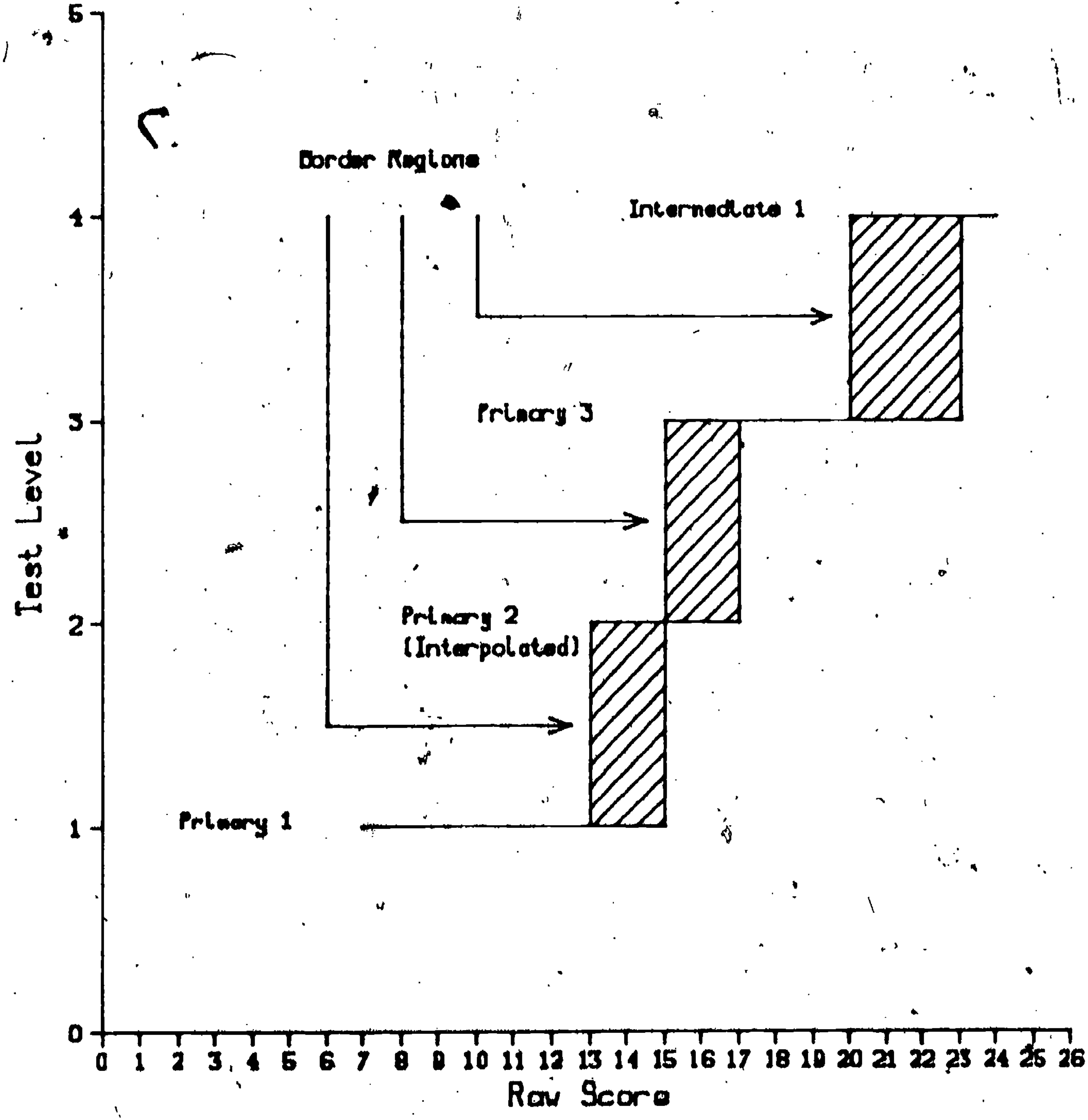


Figure 8

Interquartile Ranges of Form M2A Raw Scores for Primary 3, Intermediate 1, Intermediate 2 and Advanced Criterion Groups

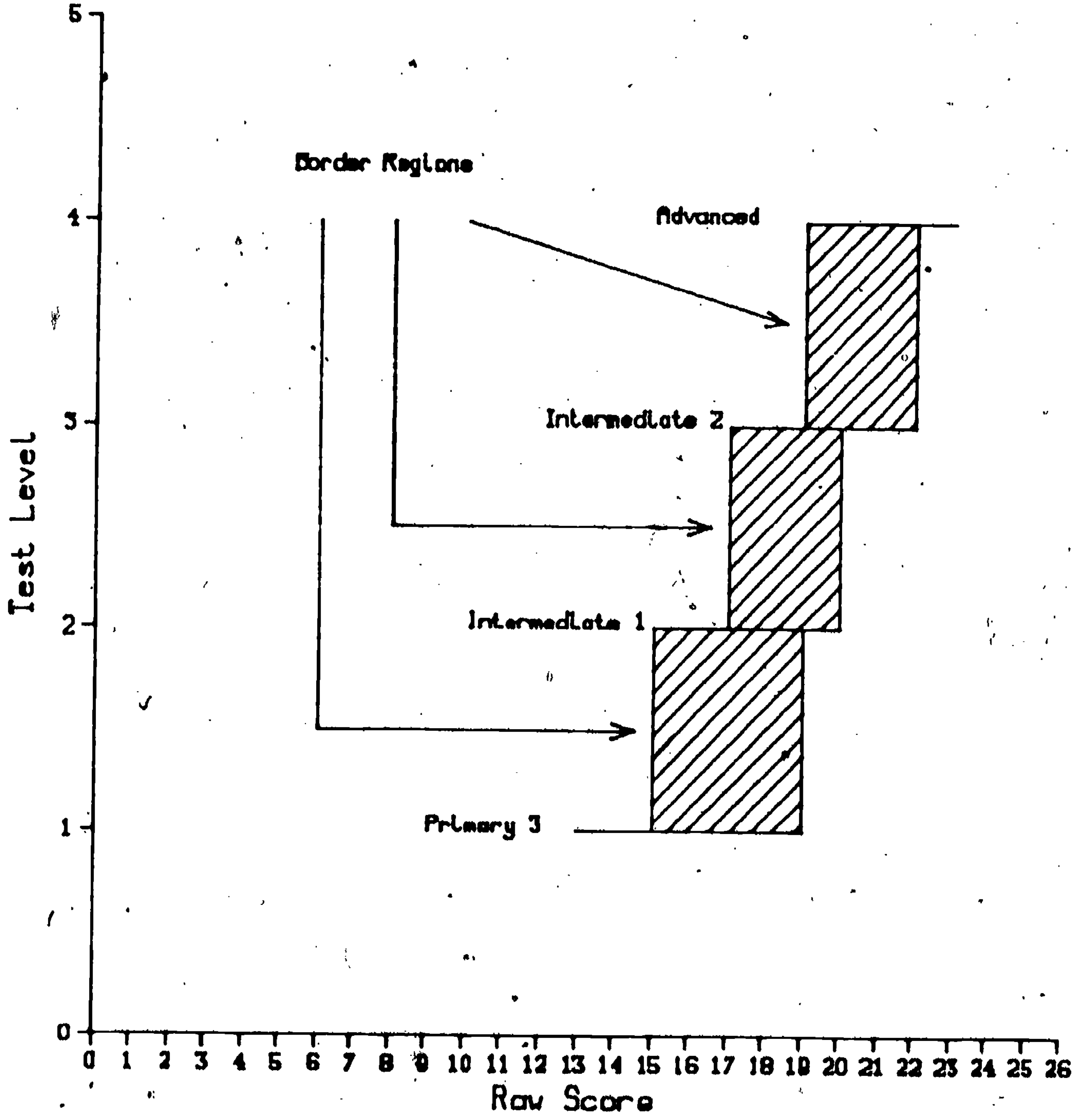
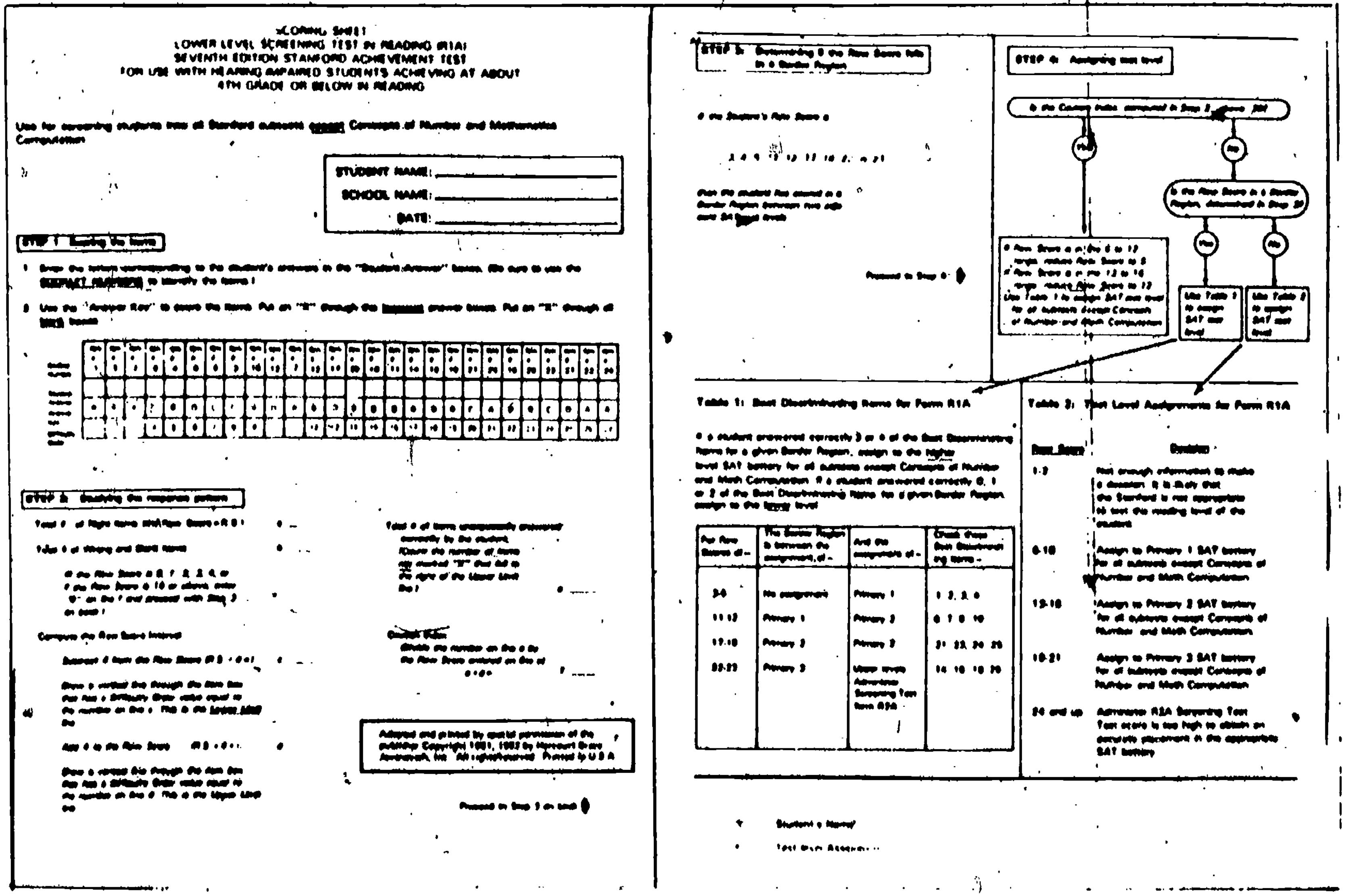


Figure 9



SCORING SHEET
LOWER LEVEL SCREENING TEST IN READING (R1A)
SEVENTH EDITION STANFORD ACHIEVEMENT TEST
FOR USE WITH HEARING IMPAIRED STUDENTS ACHIEVING AT ABOUT
4TH GRADE OR BELOW IN READING

Use for screening students into all Stanford subtests except Concepts of Number and Mathematics Computation

STUDENT NAME: ____________

SCHOOL NAME: ____________

DATE: ____________

STEP 1 Scoring the items

1 [illegible] "Student Answer" boxes. (Be sure to use the [illegible] to identify the items.)

2 Use the "Answer Key" to score the items. Put an "X" through the [illegible] answer boxes. Put an "X" through all blank boxes.

[illegible]

STEP 2 Studying the response pattern

Total # of Right items (the Raw Score = R.S.) # ___

Total # of Wrong and Blank items # ___

If the Raw Score is 0, 1, 2, 3, 4, or if the Raw Score is 10 or above, enter "0" on line [illegible] and proceed with Step 3 on back.

Compute the Raw Score Interval

Subtract 4 from the Raw Score (R.S. - 4 =) # ___

Draw a vertical line through the item box that has a Difficulty Order value equal to the number on line [illegible]. This is the Lower Limit line.

Add 4 to the Raw Score (R.S. + 4 =) #

Draw a vertical line through the item box that has a Difficulty Order value equal to the number on line #. This is the Upper Limit line.

Total # of items unexpectedly answered correctly by the student. (Count the number of items not marked "X" that fall to the right of the Upper Limit line.) # ___

Caution Index:
Divide the number on line # by the Raw Score entered on line #
÷ # = ___

Adapted and printed by special permission of the publisher. Copyright 1981, 1982 by Harcourt Brace Jovanovich, Inc. All rights reserved. Printed in U.S.A.

Proceed to Step 3 on back

STEP 3 Determining if the Raw Score falls in a Border Region

If the Student's Raw Score is

3, 4, 5, [illegible] 12, 17, 18, [illegible] 21

then the student has scored in a Border Region between two adjacent SAT levels.

Proceed to Step 4

STEP 4 Assigning test level

Is the Caution Index, computed in Step 2, above .30?

Yes

No

If Raw Score is in the 6 to 11 range, reduce Raw Score to 5.
If Raw Score is in the 12 to 16 range, reduce Raw Score to 11.
Use Table 1 to assign SAT test level for all subtests except Concepts of Number and Math Computation.

Is the Raw Score in a Border Region, determined in Step 3?

Yes

No

Use Table 1 to assign SAT test level

Use Table 2 to assign SAT test level

Table 1: Best Discriminating Items for Form R1A

If a student answered correctly 3 or 4 of the Best Discriminating Items for a given Border Region, assign to the higher level SAT battery for all subtests except Concepts of Number and Math Computation. If a student answered correctly 0, 1 or 2 of the Best Discriminating Items for a given Border Region, assign to the lower level.

For Raw Scores of –	The Border Region is between the assignment of –	And the assignment of –	Check these Best Discriminating Items –
3-5	No assignment	Primary 1	1, 2, 3, 4
11-12	Primary 1	Primary 2	6 7 8 10
17-18	Primary 2	Primary 3	21 23, 24 25
22-23	Primary 3	Upper levels Administer Screening Test form R2A	14 16 18 20

Table 2: Test Level Assignments for Form R1A

Raw Score	Decision
1-2	Not enough information to make a decision. It is likely that the Stanford is not appropriate to test the reading level of the student.
6-10	Assign to Primary 1 SAT battery for all subtests except Concepts of Number and Math Computation
13-16	Assign to Primary 2 SAT battery for all subtests except Concepts of Number and Math Computation
19-21	Assign to Primary 3 SAT battery for all subtests except Concepts of Number and Math Computation
24 and up	Administer R2A Screening Test. Test score is too high to obtain an accurate placement in the appropriate SAT battery

* Student's Name/

* Test [illegible]

BEST COPY AVAILABLE