DOCUMENT RESUME

ED 244 963 TM 840 123

AUTHOR	Livingston, Samuel A.; Zieky, Michael J.
TITLE	A Comparative Study of Standard-Setting Methods.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-83-38
PUB DATE	Oct 83
NOTE	57p.
AVAILABLE FROM	Educational Testing Service, Research Publications R116, Princeton, NJ 08541.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Achievement Tests; Comparative Analysis; *Cutting Scores; Elementary Secondary Education; *Evaluation Methods; Measurement Techniques; Scoring; *Testing; *Test Results
IDENTIFIERS	*Angoff Methods; Borderline Group Method; Comparability; Contrasting Groups Method; *Nedelsky Method

ABSTRACT

Four different systematic methods for selecting passing scores which differ primarily in the types of judgment they require were compared. The borderline group method and the contrasting groups method were each compared with the Nedelsky method at four schools and the Angoff method at another four schools. The Basic Skills Assessment Tests in reading and mathematics were administered for the study. The project was designed to determine whether different methods yield similar passing scores and, if not, whether the differences between them are systematic and predictable. The Nedelsky and Angoff methods are based upon judgments about test items. The borderline and contrasting group methods produce similar results when approximately equal numbers of students are classified as masters and non-masters. The contrasting group passing score produced different results when the ratio of masters to non-masters fluctuated. The Nedelsky and Angoff methods produced inconsistent results across schools. The passing scores were higher at schools with more able students. Results of the study suggest that those who set passing scores should use methods based upon test scores of actual test takers whenever possible. (Author/DWH)

RESEARCH REPORT

A COMPARATIVE STUDY OF STANDARD-SETTING METHODS

Samuel A. Livingston
and
Michael J. Zieky

October 1983

Educational Testing Service
Princeton, New Jersey

Acknowledgments

Our sincere thanks to the teachers, principals, and administrators at the eight participating schools for their cooperation and effort which made this research possible. Our thanks also to Phyllis Murphy and Rosemarie Arcieri, who enabled us to find these eight schools; to Edwin O. Blew and Karen Zeis Carroll, who performed the computer operations for analyzing the data; and to Georgiana Thurston and Karen Damiano, who prepared the manuscript.

Abstract

The borderline-group method and the contrasting-groups method were each compared with Nedelsky's method at four schools and with Angoff's method at another four schools, using tests of basic skills in reading and mathematic The borderline-group and contrasting-groups methods produced similar result when approximately equal numbers of students were classified as masters and nonmasters. The contrasting-groups passing score was lower than the borderline-group passing score when masters greatly outnumbered nonmasters; higher when nonmasters outnumbered masters. Results involving the Nedelsky and Angoff methods were not consistent across schools. Passing scores tend to be higher at schools where students were more able.

A Comparative Study of Standard-Setting Methods

A passing score on a test represents an answer to the question, "How much is enough?" The passing score indicates the level of knowledge or skill that will be considered sufficient for some purpose. Any method of choosing the passing score requires judgment at some stage of the process. Several different systematic methods for choosing passing scores have been suggested. (See Shepard, 1980, for a review.) These methods differ primarily in the kinds of judgment they require.

The purpose of the present study was to compare four of these methods for choosing a passing score, in an attempt to answer the following questions:

1. When these different methods are applied to the same test, with the same persons making the judgments, do they yield similar passing scores?
2. To the extent that the passing scores differ from one method to another, are these differences systematic and predictable?

The four methods investigated were "Nedelsky's method" (Nedelsky, 1954), "Angoff's method" (Angoff, 1971)*, the "borderline group method", and the "contrasting-groups method". (For a detailed description of these methods, see Livingston and Zieky, 1982.) The first two of these methods are based on judgments about the questions on the test. The judges are asked to envision the way a "borderline" test-taker would respond to each question on the test. A borderline test-taker is one whose level of the knowledge or skills measured by the test is on the borderline between sufficient and insufficient.

In Nedelsky's method (which can be used only with multiple-choice tests), the judges are asked to decide which of the wrong answer-choices the borderline test-taker could identify as not being the correct answer. These

*Angoff (personal communication, 1983) attributes this method to Ledyard R Tucker.

judgments are used to estimate, for each test question, the probability that a borderline test-taker would choose the correct answer. These probabilities are then summed, to yield the expected score for a borderline test-taker - a reasonable choice for the passing score. Angoff's method is similar to Nedelsky's, except that the judges are asked to specify the probabilities directly.

The borderline-group method follows the same logic as Nedelsky's and Angoff's methods. However, instead of making judgments about each question, the judges nominate specific individual test-takers as having a "borderline" level of the knowledge or skills the test measures. The score that is typical of these "borderline" students' performance on the test - usually the median - is taken as the passing score.

The contrasting-group method is a bit more complex. The judges classify individual test-takers as "masters" (i.e., those with a sufficient level of knowledge or skill) or "nonmasters" (i.e., those with an insufficient level of knowledge or skill). The passing score is usually chosen to minimize the number of wrong decisions (i.e., failing a "master" or passing a "nonmaster"). The passing score can also be chosen to minimize a weighted sum of the two types of wrong decisions. However, in this study, we have used the passing score that weights the two types of wrong decisions equally.

Although several previous studies have compared passing scores produced by different standard-setting methods, only a few have compared methods based on judgments of items with methods based on judgments of actual test-takers. Koffler (1980) compared Nedelsky's method with the contrasting-groups method,

for eight different tests, with results that varied considerably from one test to another. Poggio, Glasnapp, and Eros (1981) found systematic differences between methods: Ebel's method produced the highest passing score, followed by Angoff's method, the contrasting-group method, and Nedelsky's method in that order. Mills and Barr (1983) found that both Angoff's method and Ebel's method consistently produced higher passing scores than did the contrasting-groups method. Taken together, these previous findings suggest that there may be systematic differences between methods. If so, the results of the present study should reflect those differences.

Method

The tests for this study were the Basic Skills Assessment Tests in reading and mathematics, developed by Educational Testing Service. Both tests are made up of four-option multiple-choice questions. The reading test contains 65 questions; the math test contains 70. These tests are intended to test the basic reading and math skills required in the daily life of an American adult. For example, the reading test includes excerpts from a medicine bottle label, a newspaper want-ad section, a road map, and the yellow pages of a telephone directory. The math test includes problems in the four basic arithmetic operations and applications such as comparing unit prices, adding sales tax to a restaurant check, etc. "Mastery" was defined, for the purpose of this study, as the ability to perform adequately the reading/mathematical tasks of adult life in modern American society. These tasks were not specified or enumerated.

The judges for the study were teachers of students in grades 6, 7, and 8 We chose these grade levels in the hope of finding both a substantial number of students who had achieved mastery and a substantial number who had not. The judges for the reading test were teachers of English, reading, or langua arts. The judges for the math test were math teachers. (In one school, two science teachers also served as judges for the math test.)

Eight schools participated in the study, each one from a different schoo district.* These eight school districts represented a wide range of socioeconomic conditions. In each school, from three to five teachers serve as judges for each test. The schools included various combinations of grade levels (e.g., K-8, 6-8, 7-12).

The experimental design called for the teachers in each school to make judgments for three standard-setting methods: the borderline group method, the contrasting groups method, and either Nedelsky's method or Angoff's method. In four schools the teachers made judgments of their students befor making the Angoff/Nedelsky judgments; in the other four schools the order wa reversed. The resulting design, including the grade levels of the students participating, is shown in Table 1.

In the schools where the teachers judged the questions first, the researchers met with the teachers only once, for approximately two hours.

---

*One of the eight schools was, in fact, two schools, located in the same complex of buildings but administratively separate. These two schools participated together in the study, and the teachers from the two schools m together for the standard-setting sessions. In this report they will be treated as one school.

They began by explaining the purpose of the study. Next they explained the way in which individual students were to be judged as "masters", "nonmasters", or "borderline", emphasizing that these judgments were to be based on the reading or math skills necessary to function as an adult in American society. ("If this student had to do all the reading/mathematics for his/her family, could he/she do it adequately?") The form on which the teachers recorded their judgments of the students' skills also had a "cannot judge" category for students whose level of skill the teacher was unsure of. The researchers distributed these forms and the test materials. Then one researcher led the math teachers in an Angoff or Nedelsky standard-setting session, while the other researcher did the same with the reading, English, or language arts teachers. After the meeting the teachers administered the tests to their students and returned the completed answer sheets by mail.

In the schools where the teachers judged the students first, the researchers met with the teachers twice. At the first meeting, one of the researchers explained the purpose of the study, described the procedure for judging the students, and distributed the judgment forms and test materials. The researcher asked the teachers to look through the tests to see what kinds of skills the tests measured, but to use their own ideas of the skills required in daily adult life as the basis for all judgments involved in the study. At the second meeting, approximately a week later, the researchers collected the judgment forms and conducted the Angoff or Nedelsky standard-setting session.

For the Angoff standard-setting sessions the teachers were given copies the test with the correct answers indicated and a form for recording their judgments. The leader (i.e., the researcher leading the session) began by explaining the logic of the method and reviewing the definition of mastery a of the "borderline" student. Next, the leader asked the teachers to make judgments for a number of selected questions, comparing and discussing their judgments for each question. The teachers then worked independently, making preliminary judgment for each question. After approximately half an hour, t leader polled the group for their judgments on each of the questions they ha finished judging. Whenever there was a discrepancy of 20 or more percentage points between any two teachers, the leader asked the teachers with the highest and lowest judgments to explain their reasons. All the teachers wer then given the opportunity to change their judgments if they wished to do so This procedure was repeated for the remaining questions on the test.

The procedure for the Nedelsky standard-setting sessions was similar to that for the Angoff sessions, except that the correct answers to the test questions were indicated on the form the teachers used to record their judgments. The session leader did not have a fixed rule for deciding whethe or not to ask the teachers to discuss their responses. Generally, the leade would call for discussion whenever one teacher eliminated all three wrong answers and another teacher did not, or when one teacher eliminated two of t three wrong answers and another teacher did not eliminate any.

The passing score for the borderline-group method was set at the median test score of those students classified as "borderline". In those cases whe

the borderline group contained fewer than four students, no borderline-group passing score will be reported.

The passing score for the contrasting-groups method was computed by estimating a conditional probability function: the probability that a student from the combined group of masters and nonmasters with a given test score would be classified as a master. (The estimation procedure is described briefly in the Appendix to this report.) If the contrasting-groups method works as it should, this probability will increase with the student's test score. The passing score was set at the test score for which this estimated probability was equal to .50. In those cases where either group--masters or nonmasters--contained fewer than four students, no contrasting-groups passing score will be reported. (Ability-grouping in some of the schools led to this situation for some of the teachers.) Also, in those few cases where the test scores of the "masters" were no higher than those of the "nonmasters", no contrasting-groups passing score will be reported.

The passing scores for the Nedelsky and Angoff methods were the sum of the probabilities for all test items, i.e., the expected score for a borderline test-taker, as computed from the judgments.

In computing the passing scores for each school, the data were combined across teachers. For the borderline-group method, all the students judged "borderline", were combined into a single borderline group for the school. This procedure, by giving each student equal weight, tends to give a heavier weight to teachers who placed more students in the borderline group. A similar procedure was used for the contrasting-groups method. For the Nedelsky and Angoff methods, the passing scores for the individual teachers were averaged by taking a simple mean, weighting each teacher equally.

Results

Figures 1a and 1b show the passing scores for each school, as determined by each method tried at the school. On the graph for each school the small circle represents the students' mean test score; the vertical line extends one standard deviation above and below the mean. (Table A1 in the Appendix presents this information in numerical terms.) In some schools all three methods tried at the school produced similar passing scores; in other schools the three methods produced very different results. None of the methods consistently produced results similar to those of any other method.

All four methods produced passing scores that varied considerably from school to school. The contrasting-groups method showed the largest variation, producing both the highest and lowest passing scores on the math test as well as the lowest on the reading test. The borderline-group method tended to produce higher passing scores than the other methods on the reading test but not on the math test. It is difficult to generalize about the Nedelsky and Angoff methods on the basis of only four schools. The Nedelsky method produced low passing scores for both reading and math at Schools 1, 2, and 4, but high passing scores on both tests at School 3. The Angoff method tended to produce low passing scores on the reading test but not on the math test. In general, the differences between the results of the different methods were not consistent across schools.

Table 2 shows the failure rates that would have resulted from each of the passing scores at each school. Most of the differences between methods are substantial, and some are extremely large. The math test at school 3 shows

the largest differences; the contrasting-groups passing score would have failed only eight percent of the students, while the Nedelsky passing score would have failed 91 percent.

What happens when the standard-setting methods are applied separately for each individual teacher? The result of this analysis are shown in Figures 2a-2d. The passing scores for any particular method tend to be similar for teachers of the same subject in the same school, although there are exceptions. (For the Nedelsky and Angoff methods this similarity may be partly a result of the group discussion included in the procedure.)

Is it reasonable to expect the contrasting-groups method to produce a passing score similar to those produced by any of the other methods? Nedelsky's method, Angoff's method, and the borderline-group method are all based on the idea that the passing score should be the score that is typical of "borderline" test-takers. The choice of a passing score in the contrasting groups method, as applied in this study, was based on a different rationale - that of minimizing the number of misclassifications in a particular population of students. The contrasting-groups passing score depends not only on the test scores of the masters and nonmasters, but also on the number of students classified into each group. Where most of the students are masters, the masters may outnumber the nonmasters even at very low test score levels. As a result, the passing score will tend to be low (as compared with the passing scores set by other methods). Where most of the students are nonmasters, the passing score will tend to be high.

What happens when we look only at the schools in which the numbers of masters and nonmasters were approximately equal? There were six cases in which the size of the smaller group (masters or nonmasters) was at least 75 percent of the size of the larger: the reading test in Schools 1, 3, and 6, and the math test in Schools 5, 6, and 7. In one of these cases no contrasting-groups passing score could be computed. In the remaining five cases, the borderline-group passing score and the contrasting-groups passing score tended to be close to each other. The differences between these two passing scores were 3.1 and 0.8 points on the 65-point reading test and 1.8, 5.8, and 0.6 points on the 70-point math test. It seems reasonable to conclude that the contrasting-groups method and the borderline-group method will tend to produce similar passing scores when approximately equal numbers of students are judged as masters and nonmasters.

This result is corroborated by the results in the cases where one group was much larger than the other. The masters greatly outnumbered the nonmasters (by at least 5 to 1) on the reading test in Schools 2, 4, 7, and 8. In each of these schools the contrasting-groups passing score was far below the borderline-group passing score. On the math test, the masters outnumbered the nonmasters by at least 2.5 to 1 at Schools 3, 4, and 8. In each of these schools the contrasting-groups passing score was below the borderline-group passing score (although the difference was not large in School 4). In Schools 1 and 2 the nonmasters outnumbered the masters by at least 2.5 to 1; in both these schools the contrasting-groups passing score was well above the borderline-group passing score.

What is the relationship between the passing scores produced by each method and the students' ability? Table 3 shows the correlations between the passing scores and the students' mean test score. The correlations have been computed for schools and for individual teachers. In general, the teachers whose students were more able tended to set higher passing scores. The single exception was the case of the Nedelsky passing scores for the reading test, which correlated negatively with the mean scores of the teachers' students. In all other cases the correlations were positive, and, in most cases, quite large. Even the contrasting-groups method, which tends to produce a low passing score when most students are judged masters, produced higher passing scores where the students were more able.

How closely did the teachers' judgments correspond to the students' test scores? Figures 3a and 3b show the means and standard deviations of the test scores in the groups of students judged "master", "borderline", and "nonmaster" at each school. The vertical bar for each group extends from one standard deviation above the group mean to one standard deviation below. The horizontal line in the center of the bar indicates the mean. The number of students in each group is shown just below the bar. (The same information is presented in numerical form in Appendix Tables A2a and A2b, along with the means and standard deviations of scores for all participating students in each school.)

The results vary considerably from one school to another. The reading scores are shown in Figure 3a. They follow the expected pattern, with reasonably good separation, in Schools 1, 2, 4, and 6. In these schools the teachers were fairly accurate judges of their students' reading ability. In

Schools 7 and 8 the "masters" clearly scored higher than the other two groups, but the "nonmasters" scored as high or nearly as high as the "borderline" students. In School 5 the "borderline" students scored much lower than the "nonmasters". In School 3 there were no "borderline" students, and the scores of the "masters" differed very little from those of the "nonmasters".

The math scores are shown in Figure 3b. The scores of the three groups - "masters", "borderline", and "nonmasters" - follow the expected pattern in all eight schools, with reasonably good separation of the groups in most of the schools. In Schools 1, 6, and 7 there was a large overlap between the scores of the "borderline" students and those of the "nonmasters", and in School 3 all the differences between groups were small. Schools 4 and 5 provide an interesting comparison. Although the "nonmasters" in School 4 scored higher than the "masters" in School 5, the teachers' judgments of their students in each of these schools corresponded quite well to the students' test scores (which were not available to the teachers at the time they made their judgments).

Figures 4a and 4b show the mean scores for students judged "master", "borderline", and "nonmaster" by each teacher. Only the means based on four or more students are shown. With only one exception, the order of the three group mean scores for each teacher is as it should be: "masters" highest; "nonmasters" lowest. However, in some cases there was very little separation between group means for the same teacher. For example, a teacher's "borderline" students may have scored only slightly higher than the same teacher's "nonmasters". The group means seem to imply some striking differences in

standards between teachers in the same school. For example, on the reading test in School 3, one teacher's "nonmasters" scored higher than the other two teachers "masters". A similar situation occurred for the math test in School 1, where one teacher's "masters" scored about as low as the other two teachers' "nonmasters". More commonly, one teacher's "borderline" students would score as high as another teacher's "masters" or as low as another teacher's "nonmasters".

The information in Figures 3a and 3b provides a good indication of the extent to which the borderline-group method was working when applied to the full group of students classified as "borderline" in each school. Ideally, the test scores of the borderline group should have a small standard deviation and a mean score between the means for the masters and the nonmasters. In general, the standard deviation of scores for the borderline group was large - nearly as large as for the full group of students participating in the study. On the average (across schools), the borderline group standard deviation was 87 percent of the total-group standard deviation for the reading test and 86 percent for the math test. However, in 12 of the 15 cases, the borderline group mean score was clearly between the mean scores for masters and nonmasters. The exceptions were Schools 5 and 8 for the reading test and School 6 for the math test, where the scores of the borderline group were as low as (or lower than) those of the nonmasters. (The borderline-group method could not be applied to the reading test in School 3, where no students were classified as "borderline" in reading.)

One possible reason for the large variation in the scores of the borderline group might be differences between individual teachers. If so, the borderline-group method might work better when applied separately to each individual teacher's own borderline group. Figures 4a and 4b show that for most teachers, the mean score of the borderline group is clearly between the mean scores for the masters and the nonmasters. However, Figures 5a and 5b show that, even for individual teachers, the standard deviation of the borderline group tends to be nearly as large as the standard deviation of the scores of all the teacher's students. For the reading teachers, the standard deviation of the borderline group averages 8.0 score points, which is 83 percent as large as the average total-group standard deviation. For the math test the corresponding figures are 7.9 score points and 85 percent.

Figures 6a, 6b, 6c, and 6d are graphs of the conditional probability functions used to determine the contrasting-groups passing scores for each school.* If the method worked as it should, the graph will show a curve rising from nearly zero probability at low test score levels to nearly one at high score levels. The passing score that misclassifies the fewest students will be the score level at which the probability of being judged a master is .50. We have used the symbol "X_{50}" to refer to this score level. The expression "$X_{75} - X_{25}$" represents the difference between the test scores that correspond to a 75 percent probability and a 25 percent probability of being judged a master.

*The method used to estimate these functions is described in the Appendix. The symbols "A" and "B" at the top of each graph refer to a formula given in the Appendix.

The smaller this difference, the sharper the separation between the scores of the "masters" and "nonmasters", and the better the contrasting-groups method is working. Another measure of the extent to which the test is separating "masters" from "nonmasters" is the slope of the curve at X_{50}, which is its steepest point. The steeper the slope, the better the separation. This slope is also indicated on each graph.

Figures 6a and 6b show the curves for the reading test scores, combining the judgments of the teachers in each school. The method seems to have worked fairly well in Schools 1, 2, 4 and 6; not as well in Schools 7 and 8; poorly in School 5; and not at all in School 3. (The problem in School 3 seems to be the differences between individual teachers' standards, as can be seen in Figure 4a.) The four schools in which the method worked fairly well were those in which it yielded the highest standards.

The conditional probability curves for the math test scores, shown in Figures 6c and 6d, present a very different picture. There were no schools where the method failed completely. The method worked extremely well in School 5, where the resulting standard was low, and quite well in School 4, where the standard was high. It worked nearly as well in School 3, where it produced the lowest standard, as in School 2, where it produced the highest standard. In general, the schools where the contrasting-groups method worked best were not the same for math as for reading.

Table 4 shows the number and the percentage of correct and incorrect classifications resulting from the use of the contrasting-groups passing score in each school. (In this case, "correct" means "the same as the teacher's

classification of the student.") Notice that where the overwhelming majority of the students are masters, the use of the contrasting-groups passing score will tend to misclassify most of the nonmasters (unless the separation between the groups is nearly perfect). This situation occurs for the reading test in Schools 2, 7, and 8, and, to a lesser extent, in School 4. In this case, it would be possible to obtain a fairly high percentage of correct classifications by disregarding the test scores and passing all the students. It is much harder to avoid classification errors when the numbers of masters and nonmasters is nearly equal, as is the case for math in School 5.

These results suggest a question for further investigation: What would happen if the data from the contrasting-groups method were used to set a standard based on a rationale similar to that of the other methods? Such a standard might be based on the idea of classifying a student into the group - master or nonmaster - in which the student's test score would be more typical. It would not depend on the number of students classified into each group, but only on their test scores. As a result, it would not minimize the number of wrong decisions unless the test-taker population contained equal numbers of masters and nonmasters. Therefore, such a standard would not be a wise choice as a passing score, except in this special case. But would such a standard agree closely with the passing scores set by the borderline-group method?

To investigate this question, we defined a number which we call "C2", computed from the contrasting-groups data by the formula

$$C2 = \frac{\bar{x}_m(1/s_m) + \bar{x}_n(1/s_n)}{1/s_m + 1/s_n}$$

where $\bar{x}$ represents the mean score, s represents the standard deviation of scores, and m and n refer to the groups of masters and nonmasters. This number, used as the basis for classification, places a student into the group in which the student's test score would be fewer standard deviations from the mean. For example, if a student's test score would be 0.7 standard deviations below the mean in the group of "masters" but 0.8 standard deviations above the mean in the group of nonmasters, the student's score would be above the "C2" standard.

Figure 7 shows the relationship between "C2" and the borderline-group passing score. Each data point represents one school. Obviously, the two standards agree closely. Not only is the correlation very high, but 14 of the 15 data points are very close to the diagonal line y=x.

Discussion

The purpose of this study was to answer two important questions about the four standard-setting methods investigated:

1. Do the different methods yield similar passing scores?
2. If not, are the differences between methods systematic and predictable?

The comparison between the borderline-group method and the contrasting-groups method provided a qualified "yes" answer to the first question and a "yes" to the second. Where the number of "masters" and the number of "nonmasters" were similar (i.e., differed by 25 percent or less), the contrasting-groups method and the borderline-group method yielded similar passing scores. Where the

"masters" greatly outnumbered the "nonmasters", the contrasting-groups method produced a lower passing score than the borderline-group method, as could be expected. Conversely, where the "nonmasters" greatly outnumbered the "masters", the contrasting-groups method produced the higher passing score.

The comparisons involving the Nedelsky and Angoff methods were far less predictable. Each of these methods was applied at four schools, and for both methods the results varied from school to school.

The results of this study reflected a general tendency for teachers of higher-ability students to set higher standards. (The single exception was the application of the Nedelsky method to the reading test.) This finding was not an artifact of the methods. There was no suggestion to the teachers or any limit on the number of students to be classified as masters or nonmasters. Indeed, some teachers did classify all or nearly all of their students into the same group. There was no suggestion of a relative standard in the verbal definition given to the teachers. The standard was to represent the minimum level of reading/math skill necessary to function adequately as an adult in American society. Possibly the teachers at the schools with more able students envision a different type of adult life for their students than do the teachers at the schools where students are less able.

One surprising finding was the large variation in the test scores of the students classified as "borderline". This group often included some of the highest and lowest scoring students, despite the availability of a "cannot judge" category for students whose skills the teachers did not feel confident in judging. The large variation in the scores of the borderline group occurred for individual teachers as well as for schools.

What accounts for the presence of many high and low scoring students in the borderline group? More generally, why were the teachers' judgments of their students' skills often inconsistent with the students' test scores? One possibility is that the teachers may have been unaware of the skills (or lack of skills) of some of their students. The skills tested were those taught in elementary school; junior high school teachers may not observe these skills systematically. Another possibility is that the teachers may have based their judgments on reading or math skills other than those measured by the test, or on only a subset of the skills measured by the tests. The teachers might not have agreed completely with the developers of the tests as to which specific reading and math skills are necessary to function adequately as an adult. A third possibility is that the test scores may simply have been invalid for some students. Some students may not have made a genuine effort; some may have copied from their neighbors' papers. Taken together, these three explanations may account for the inconsistencies between test scores and teachers' judgments. And in most cases, despite these inconsistencies, the agreement between test scores and judgments was good enough to provide a fairly clear choice of a passing score, by either the contrasting-groups method or the borderline-group method.

Possibly the most important finding of our study is that the results of the Nedelsky and Angoff methods were generally not consistent with those of the borderline-group method. The Nedelsky and Angoff judgments did not accurately reflect the actual test performance of real students classified as "borderline". But were the results of the borderline-group method valid? In

12 of 15 cases, the borderline group mean score was clearly between the mean scores for masters and nonmasters. In the other three cases, the test scores of the borderline group were close to or below those of the nonmasters, but in two of these three cases the Angoff passing score was even lower than the borderline-group passing score. Any correction applied to the borderline-group passing score would have moved it even farther from the Angoff passing score.

This finding leads us to suggest that those who set passing scores use methods based on the test scores of real test-takers whenever possible. Those who use the Nedelsky and Angoff methods might consider a modification that allows the judges to revise their judgments on the basis of actual student response data from the test.

The results of this study might have been different if the teachers at each school had been required to agree on a precise verbal definition of the standard in behavioral terms before judging their students or the test questions. If the teachers had specified their own standard in terms of the specific reading or math tasks that distinguish "masters" from "nonmasters", the teachers' judgments of their students might have been more consistent with their judgments of the test questions. Unfortunately, the teachers' time available for this study was not sufficient to allow for such a step in the procedure. This step could be the missing link that provides for consistency between standard-setting methods based on judgments about students and methods based on judgments about test questions.

Table 1: Design of the Study

School*	Community Type	Grade Level of Students	Judgments	
			First	Second
1	Urban	6, 7, 8	Questions (Nedelsky)	Students
2	Non-urban	7, 8	Questions (Nedelsky)	Students
3	Urban	7, 8, 9	Students	Questions (Nedelsky)
4	Non-urban	8	Students	Questions (Nedelsky)
5	Urban	6, 7, 8	Questions (Angoff)	Students
6	Urban	7, 8	Questions (Angoff)	Students
7	Urban	7, 8	Students	Questions (Angoff)
8	Non-urban	7, 8	Students	Questions (Angoff)

*The numbering of the schools is arbitrary and does not correspond to the order in which data were collected.

Table 2: Failure rates resulting from passing score determined by each method in each school.

		Contrasting Groups	Borderline Group	Nedelsky	Angoff
Reading Test					
School	1	.44	.51	.31	--
	2	.05	.26	.04	--
	3	--	--	.68	--
	4	.11	.31	.05	--
	5	.30	.24	--	.11
	6	.46	.46	--	.18
	7	.03	.30	--	.30
	8	.04	.21	--	.07
Math Test					
School	1	.89	.59	39	--
	2	.82	.65	.20	--
	3	.08	.45	.91	--
	4	.28	.33	.24	--
	5	.55	.45	--	.55
	6	.64	.37	--	.77
	7	.42	.38	--	.47
	8	.20	.41	--	.76

Table 3: Correlations between passing score and students' mean test scores

Passing score	Reading Test		Math Test	
	Schools	Teachers	Schools	Teachers
Contrasting groups	.33	.43	.64	.65
Borderline group	.91	.70	.93	.84
Nedelsky	a	-.33	a	.70
Angoff	a	.61	a	.85

a Correlations based on fewer than eight observations were not computed.

NOTE: The four columns of this table correspond, respectively, to Figure 1a, Figures 2a and 2b, Figure 1b, and Figures 2c and 2d.

Table 4: Correct and incorrect classifications resulting from contrasting groups method.

		Number of students "Masters" Pass	Number of students "Masters" Fail	Number of students "Nonmasters" Pass	Number of students "Nonmasters" Fail	Proportion correctly classified "Masters"	Proportion correctly classified "Nonmasters"	Proportion correctly classified Combined	Slope of conditional probability curve
eading Test									
School	1	97	18	14	84	.84	.86	.85	1/20
	2	134	3	12	4	.98	.25	.90	1/26
	3	--	--	--	--	--	--	--	0
	4	123	6	15	10	.95	.40	.86	1/17
	5	44	3	17	4	.94	.19	.71	1/58
	6	113	17	16	86	.87	.84	.86	1/20
	7	213	5	25	4	.98	.14	.88	1/39
	8	220	6	37	4	.97	.10	.84	1/41
ath Test									
School	1	10	26	8	120	.28	.94	.79	1/34
	2	33	18	10	220	.65	.96	.90	1/21
	3	70	3	27	2	.96	.07	.71	1/27
	4	114	13	14	32	.90	.70	.84	1/18
	5	63	11	10	68	.85	.87	.86	1/12
	6	81	64	37	136	.56	.79	.68	1/39
	7	101	29	33	59	.78	.64	.72	1/29
	8	150	16	28	28	.90	.50	.80	1/31

Figure 1a. Passing scores for the reading test, computed for each school: C = Contrasting-groups; B = Borderline-group; N = Nedelsky; A = Angoff.

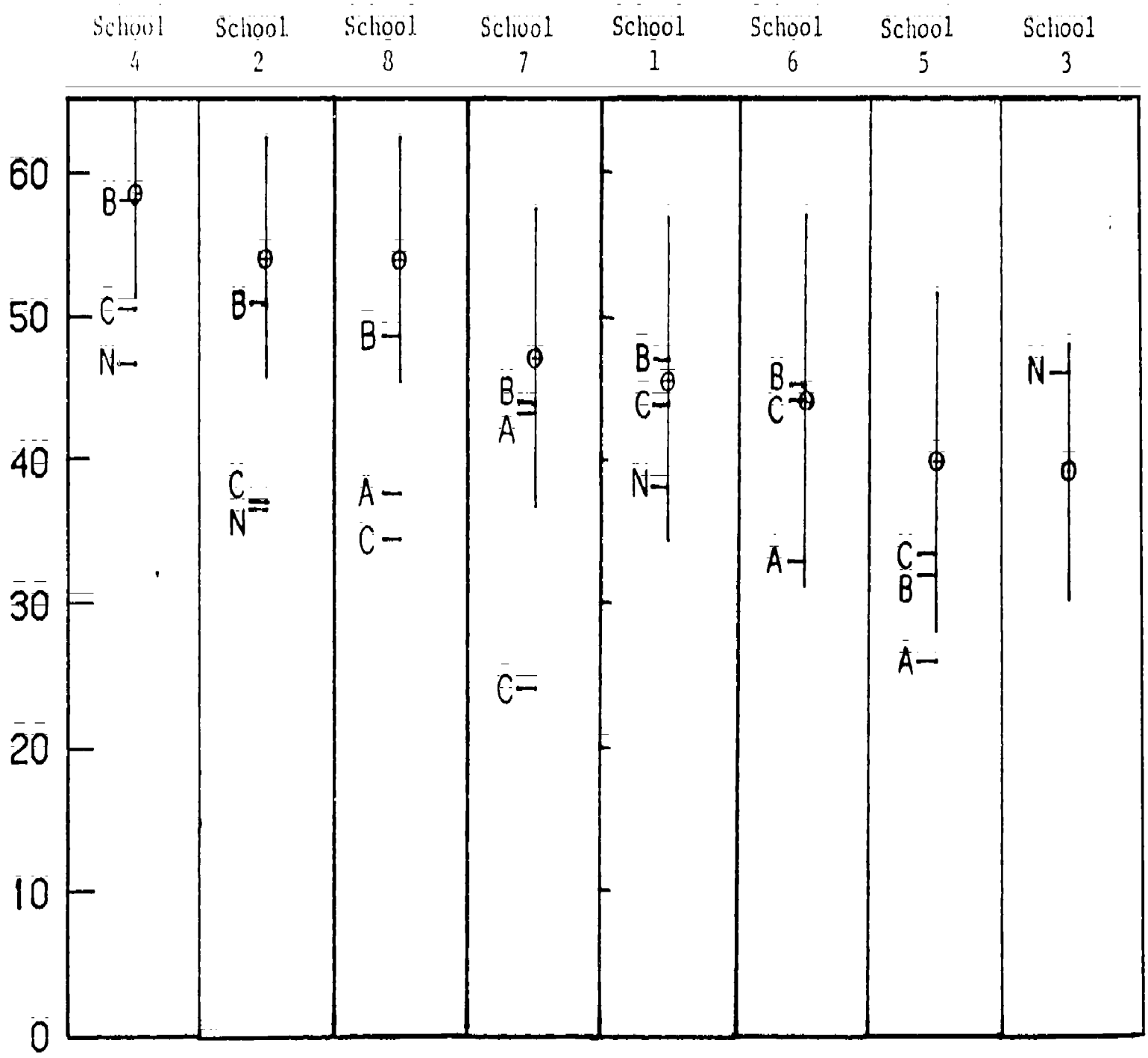


Figure 1b. Passing scores for the math test, computed for each school: C = Contrasting-groups; B = Borderline-group; N = Nedelsky; A = Angoff.

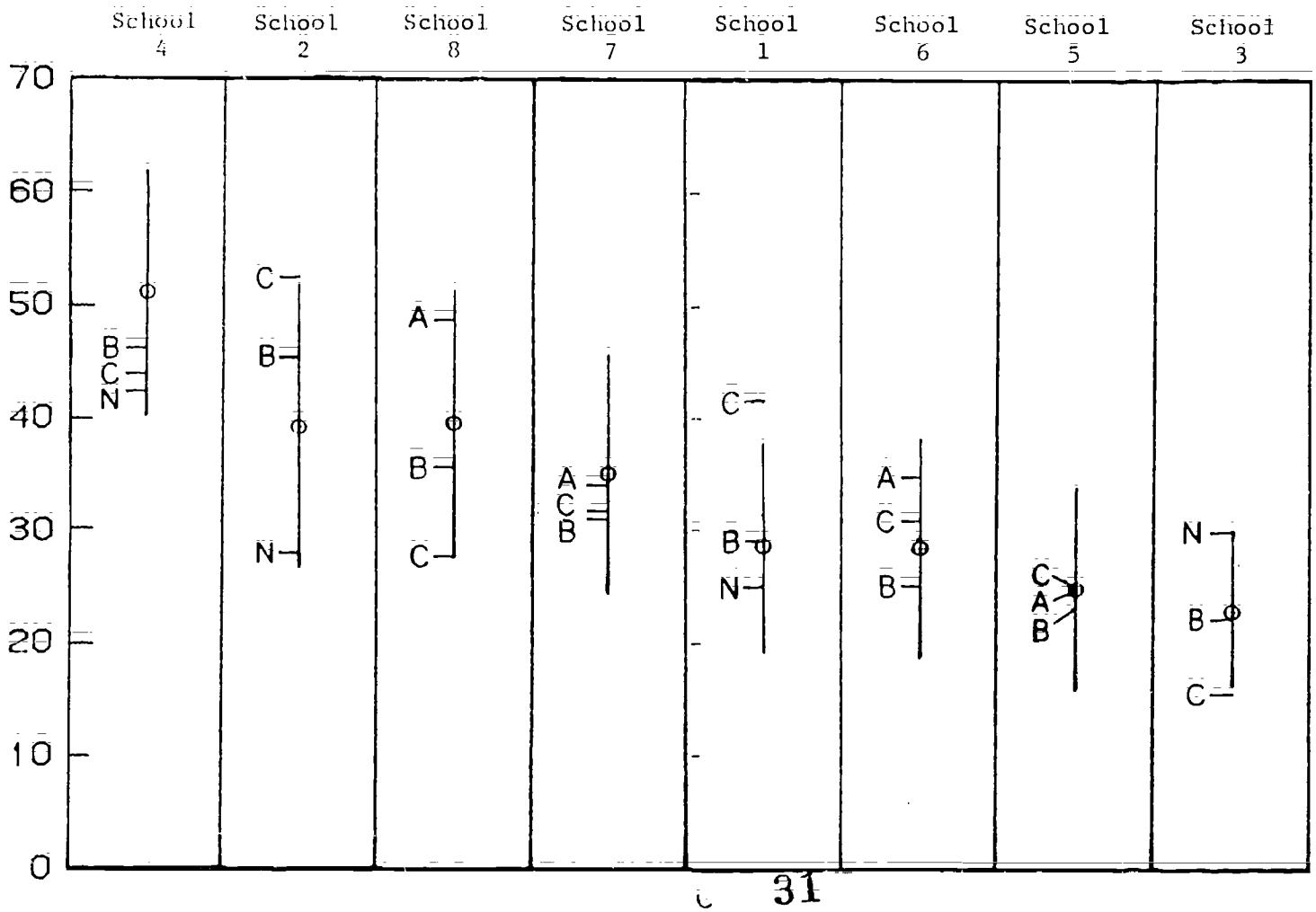


Figure 2a. Passing scores for the reading test, computed for individual teachers in Schools 1-4.
(C = Contrasting-groups; B = Borderline-group; N = Nedelsky)

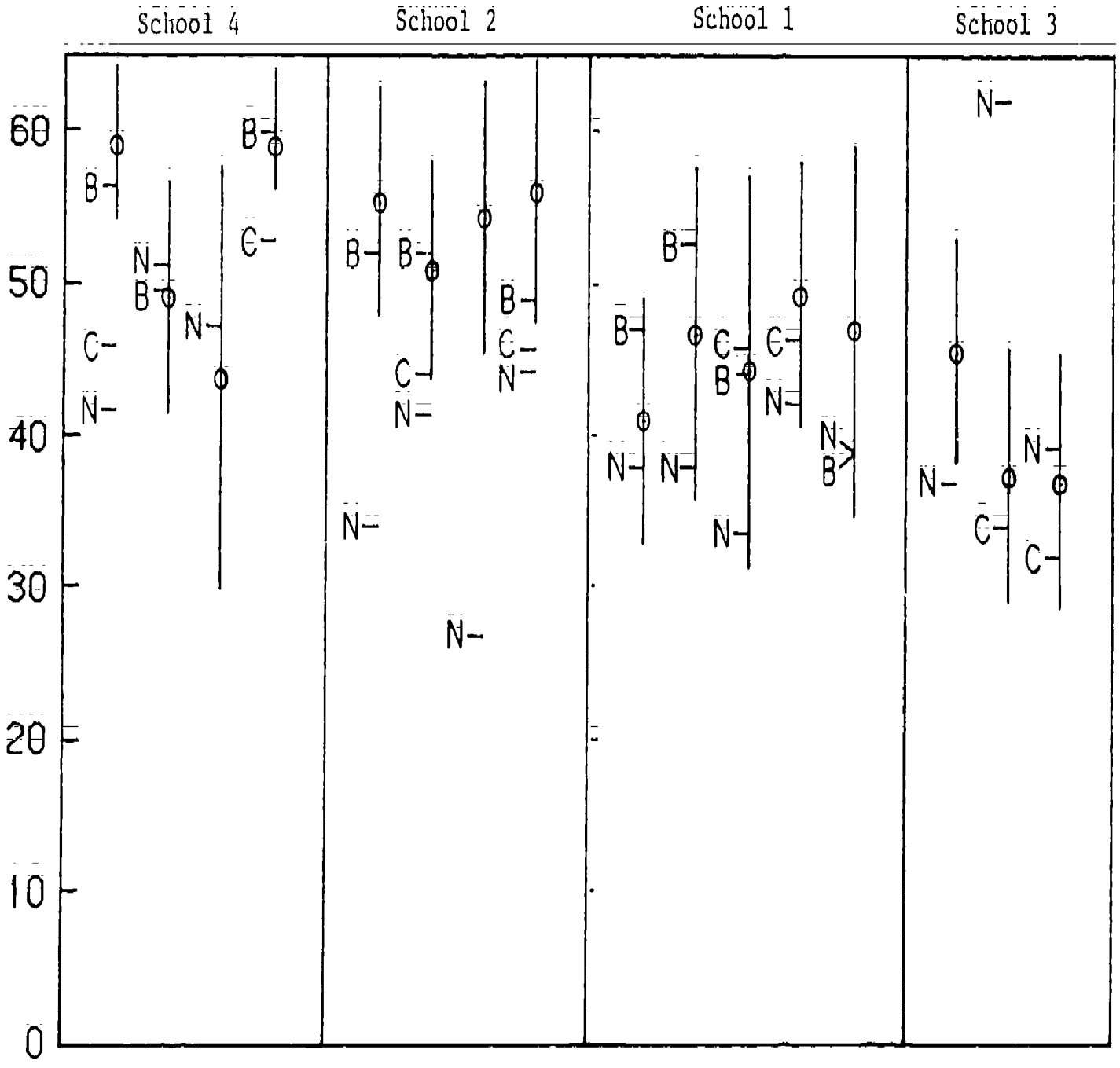


Figure 2b. Passing scores for the reading test, computed for individual teachers in Schools 5-8.
(C = Contrasting-groups; B = Borderline-group; A = Angoff)

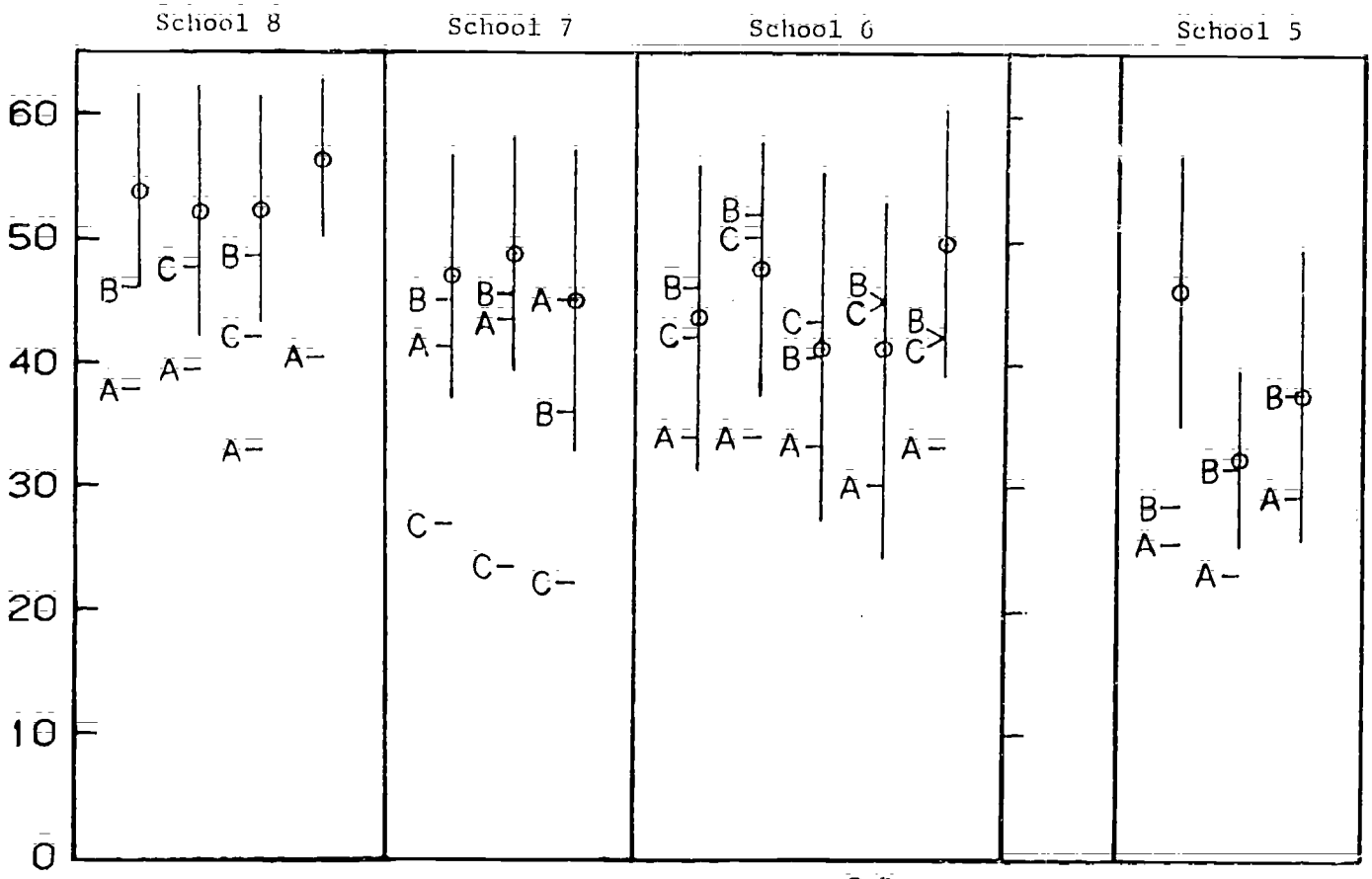


Figure 2c. Passing scores for the math test, computed for individual teachers in Schools 1-4.
(C = Contrasting-groups; B = Borderline-group; N = Nedelsky)

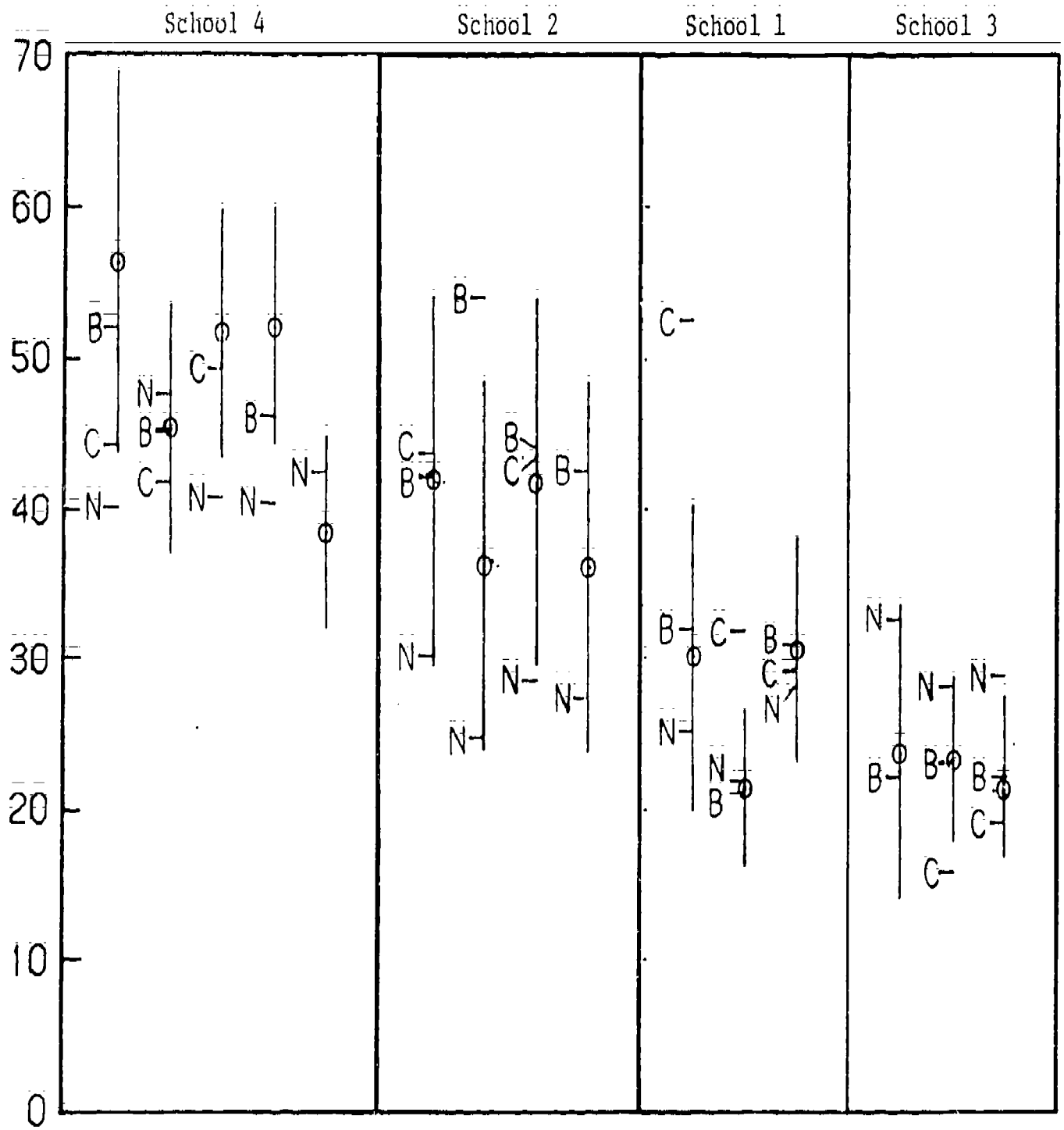


Figure 2d: Passing scores for the math test, computed for individual teachers in Schools 5-8.
(C = Contrasting-groups; B = Borderline-group; A = Angoff)

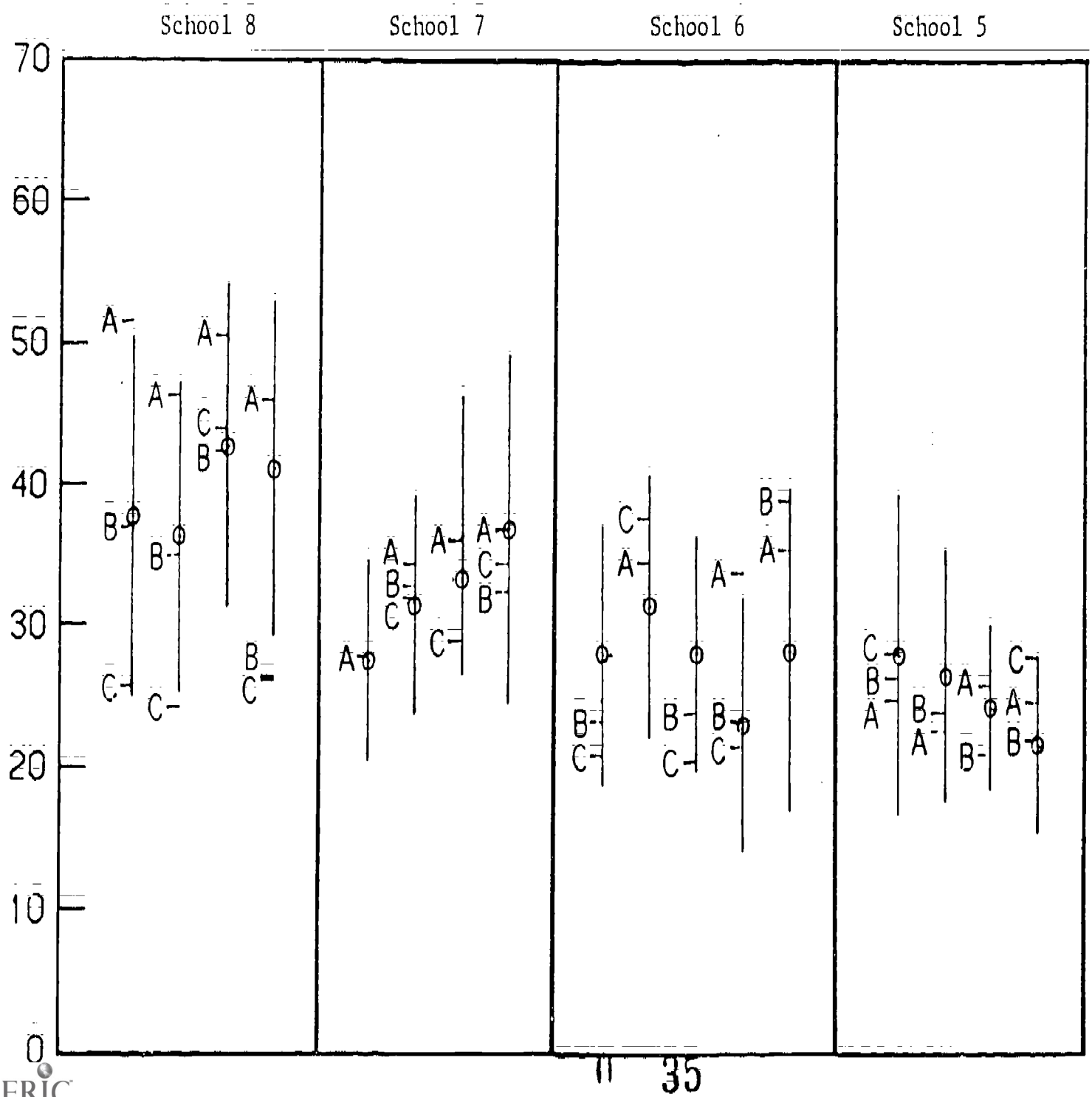


Figure 3a

Means and standard deviations of reading test scores of students judged "Master", "Borderline", and "Nonmaster" in each school.

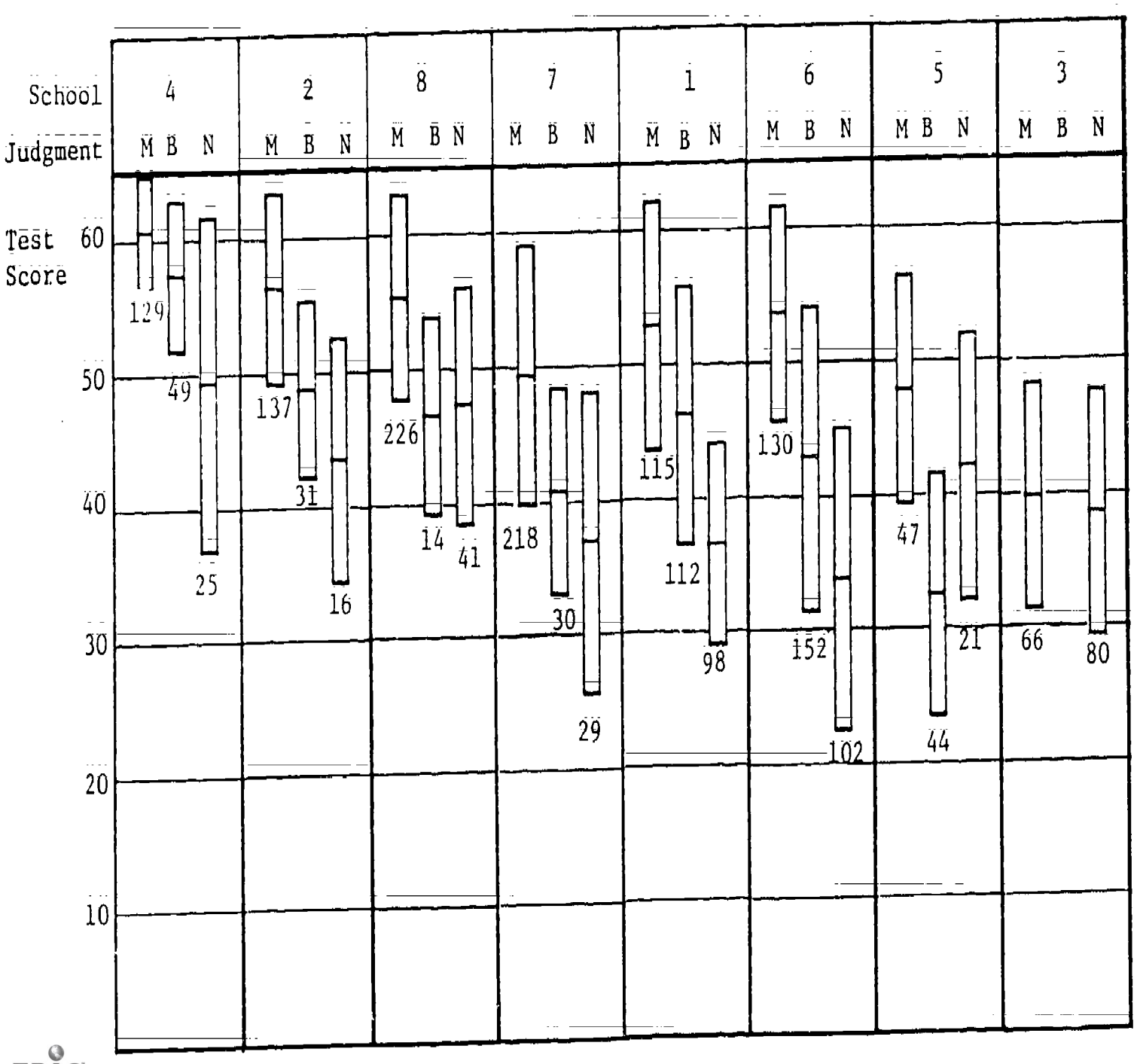


Figure 3b

Means and standard deviations of mathematics test scores of students judged "Master", "Borderline", and "Nonmaster" in each school.

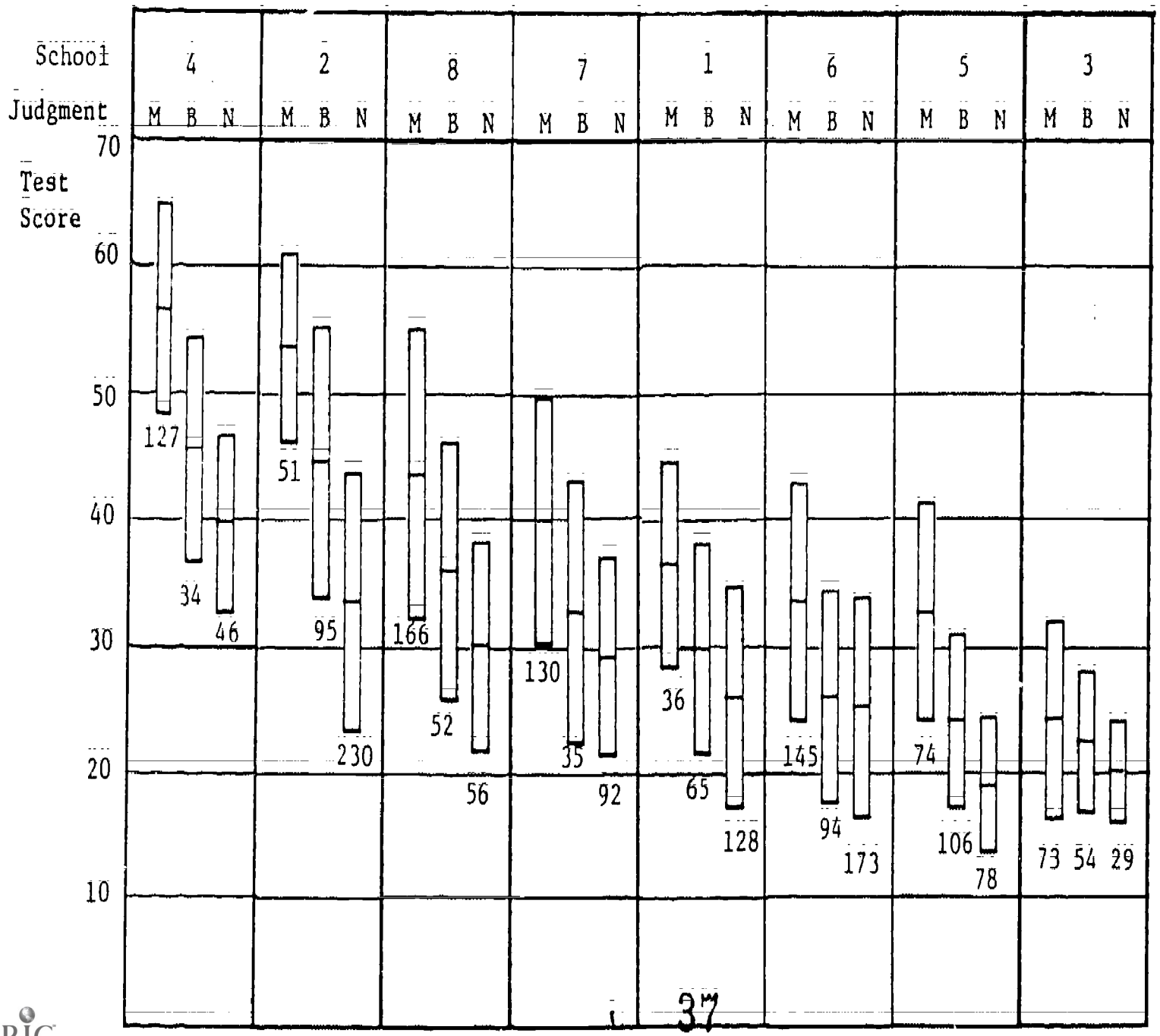


Figure 4a. Mean reading test scores of students judged "master", "borderline", and "nonmaster" by each teacher.

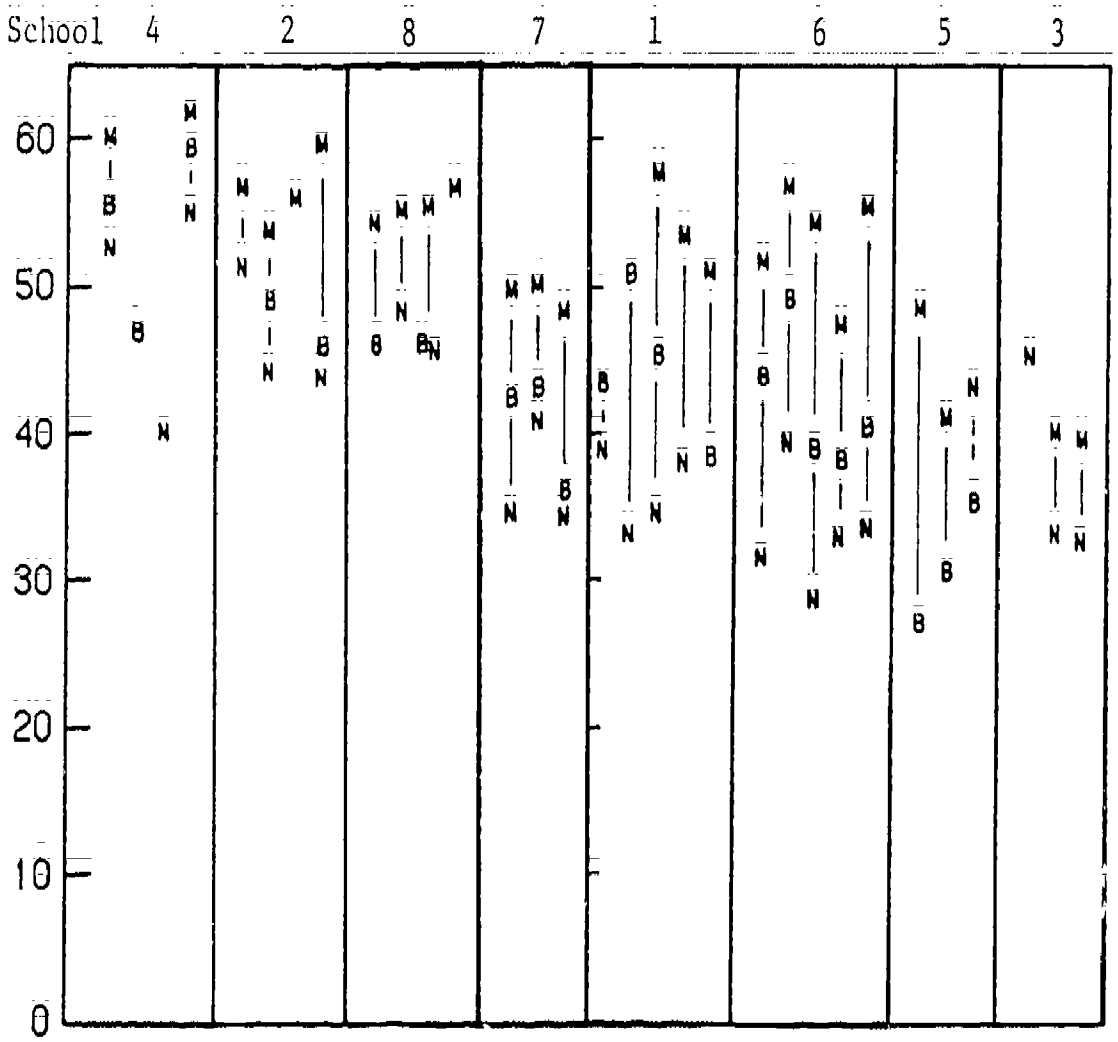


Figure 4b: Mean math test scores of students judged "master", "borderline", and "nonmaster" by each teacher.

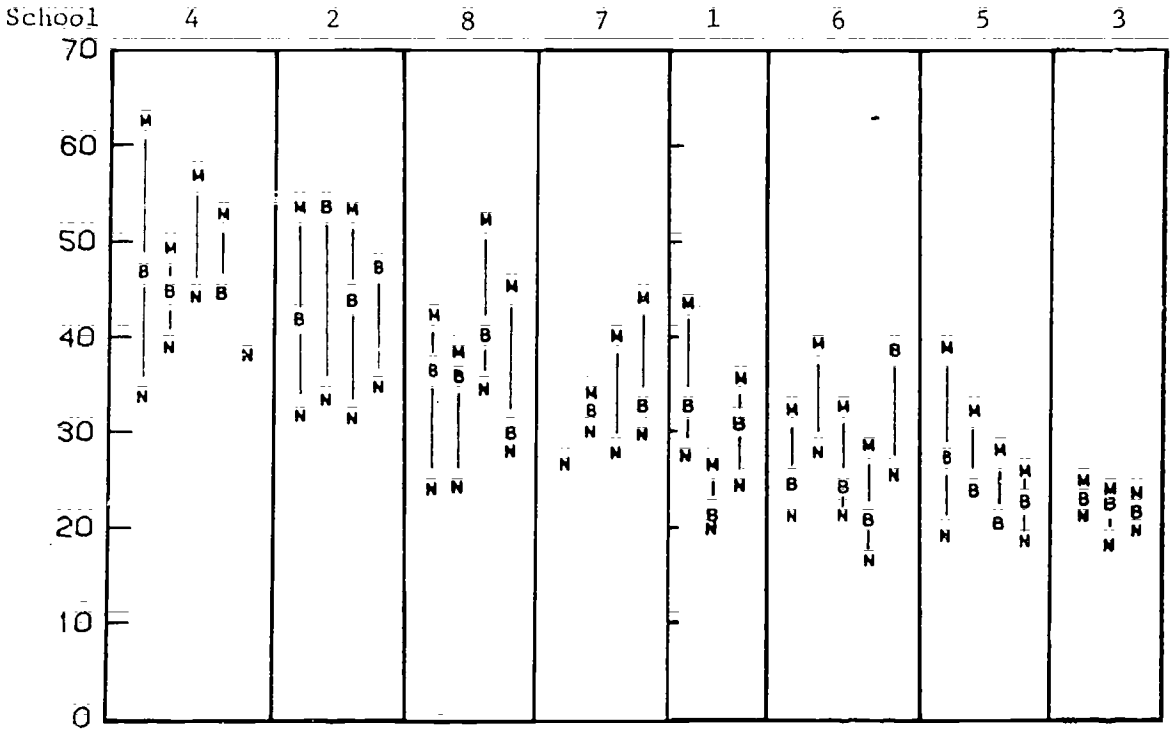


Figure 5a. Distribution of standard deviation of test scores for borderline group and for all students: reading test

Standard Deviation of Test Scores	Students Judged "Borderline"		All Students	
	Schools	Individual Teachers	Schools	Individual Teachers
0- 2				
2- 4		XX		
4- 6	X	X		XXX
6- 8	XXX	XXXXXXXX	X	XXXXXX
8-10	XX	XXXXXXX	XXX	XXXXXXXXX
10-12	X	XXXX	XXX	XXXXXX
12-14			X	XXXX
14-16		X		XXX
16-18				
18-20				
20 or more				
not computed*	X	XXXXXXXX		

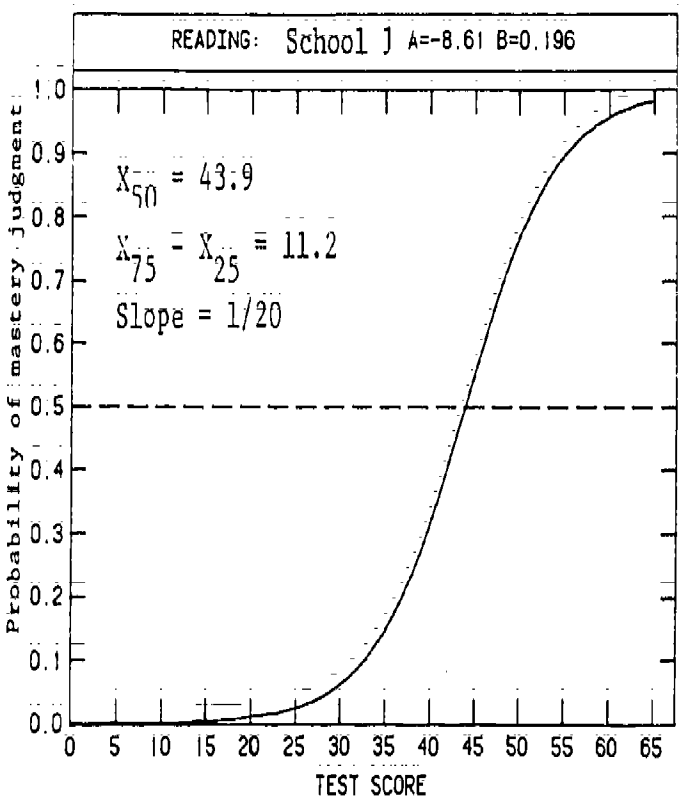
*Fewer than 4 students

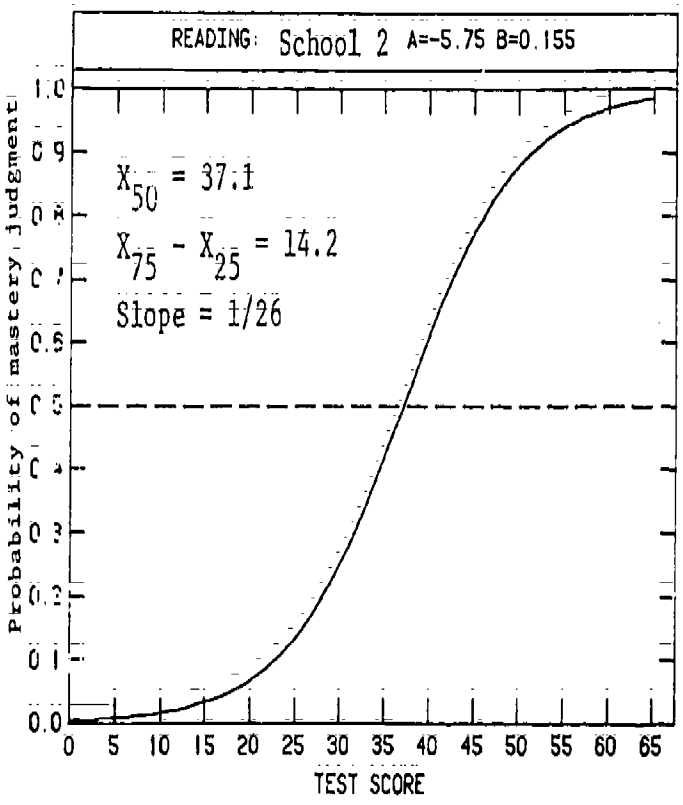
Figure 5b. Distribution of standard deviations of test scores for borderline group and for all students: math test

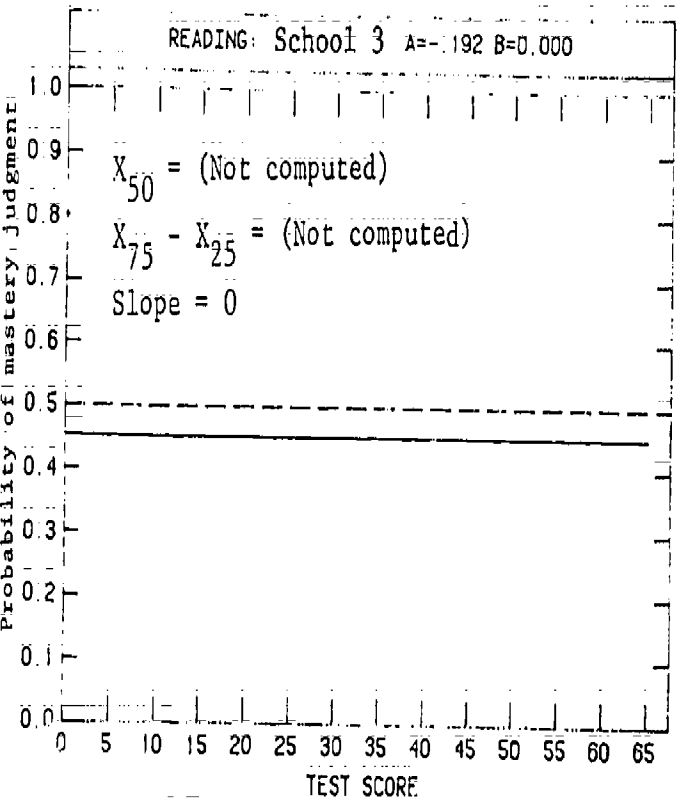
Standard Deviation of Test Scores	Students Judged "Borderline"		All Students	
	Schools	Individual Teachers	Schools	Individual Teachers
0- 2				
2- 4				
4- 6	X	XXXXXXX		XXXX
6- 8	X	XXXXXX	X	XXXXXX
8-10	XXX	XXXXXX	XXX	XXXXXXXXX
10-12	XXX	XXXXX	XXX	XXXXXX
12-14		XX	X	XXXXXXX
14-16		X		
16-18				
18-20				
20 or more				
not computed*		XXXXX		

*Fewer than 4 students

Figure 6a. Estimated relationship between reading test scores and probability of mastery judgment: Schools 1-4.







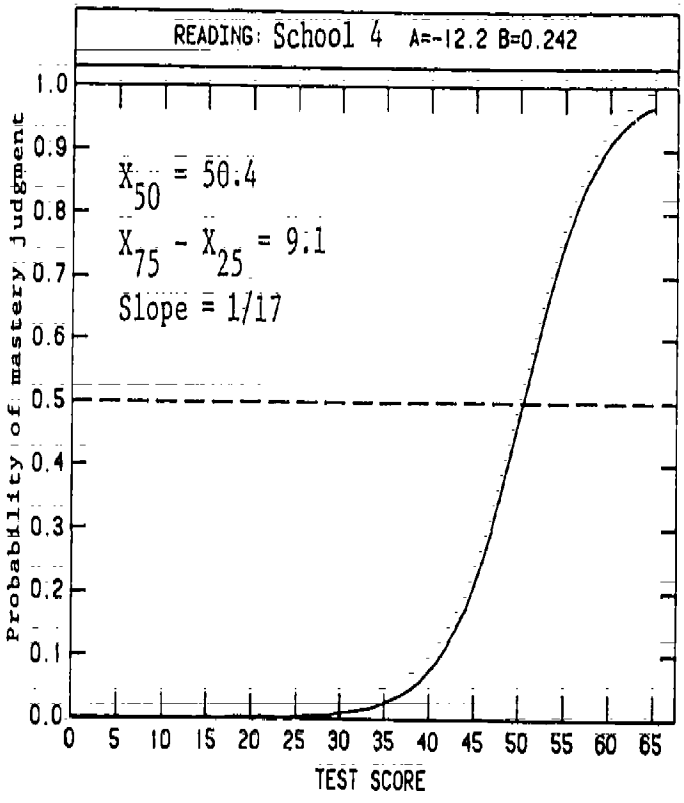
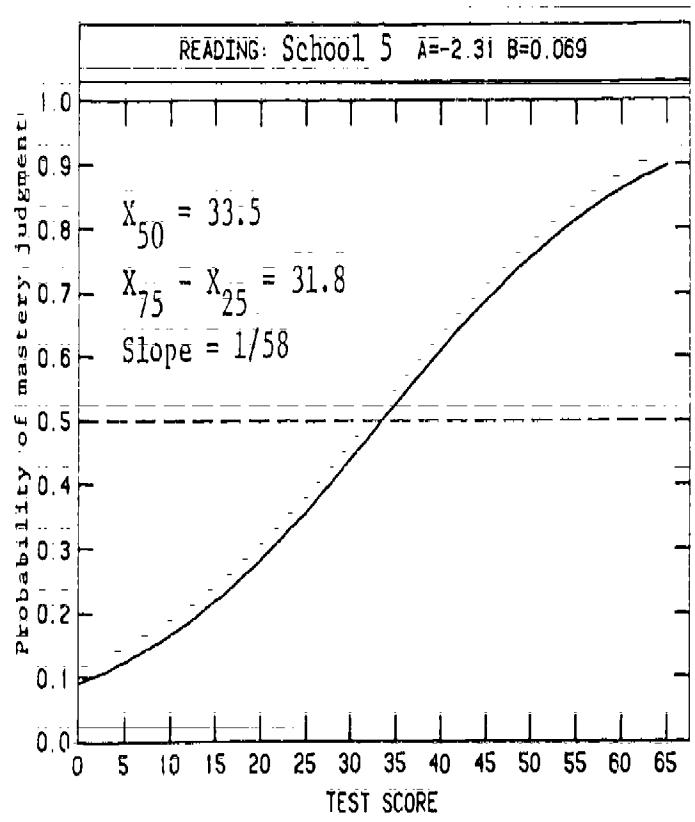
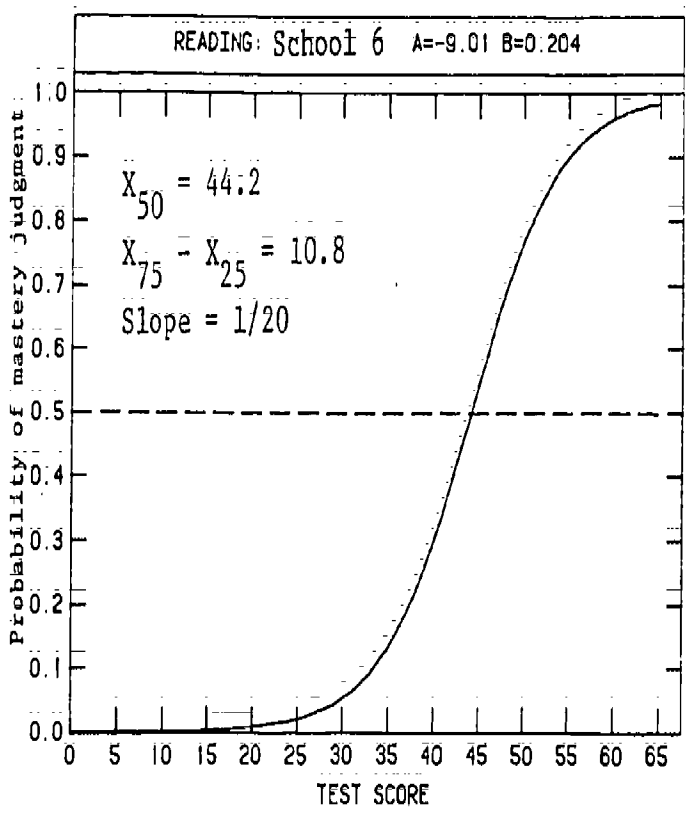
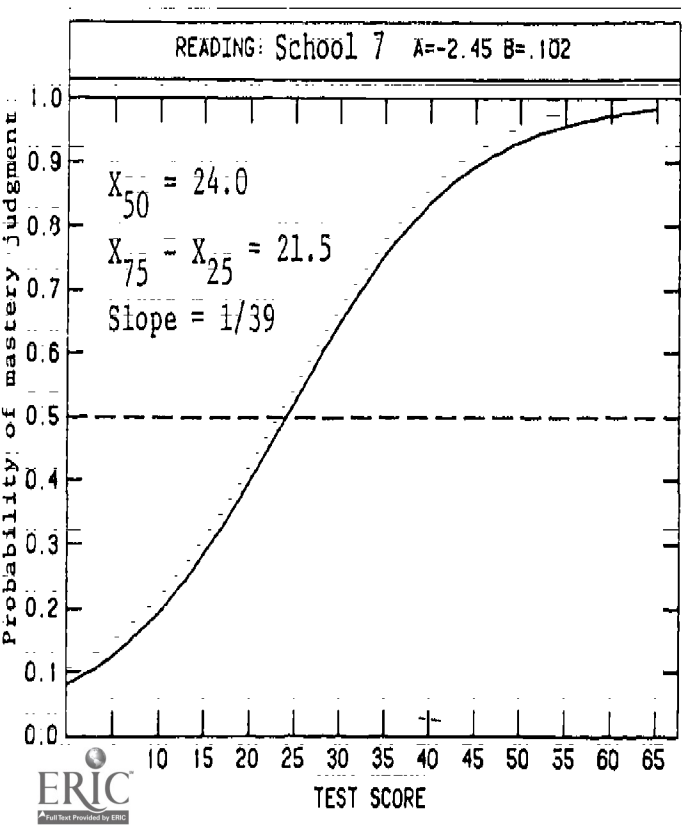


Figure 6b. Estimated relationship between reading test scores and probability of mastery judgment: Schools 5-8.







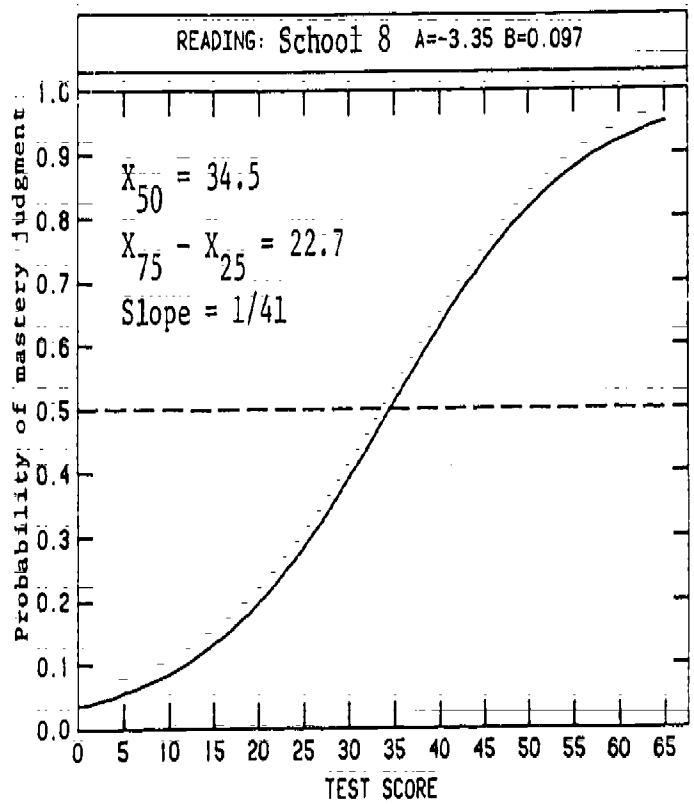
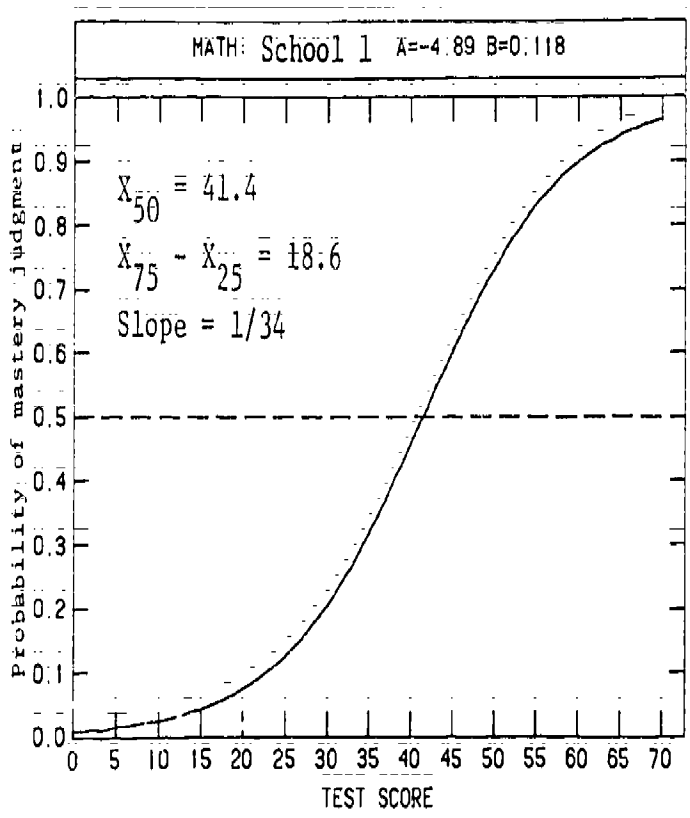
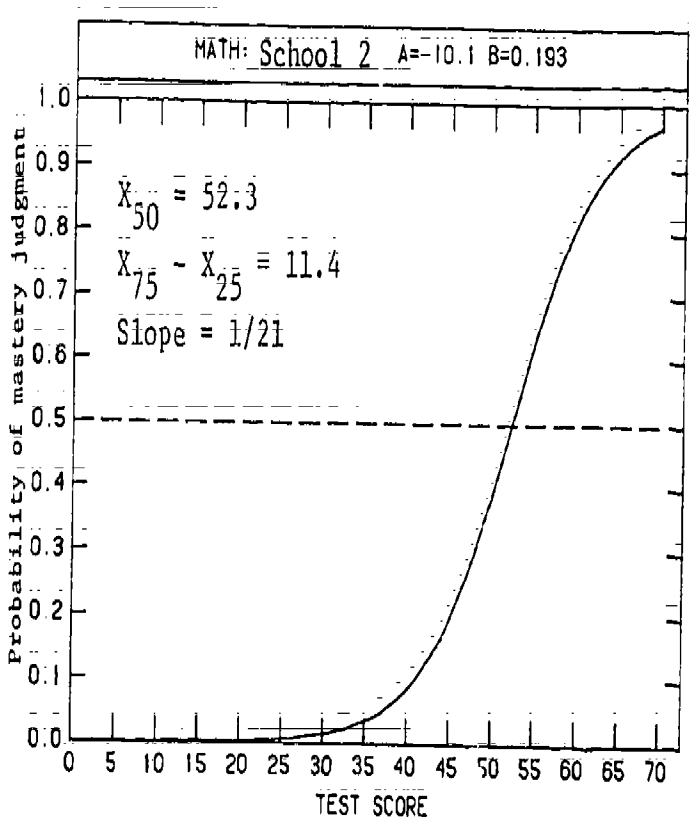
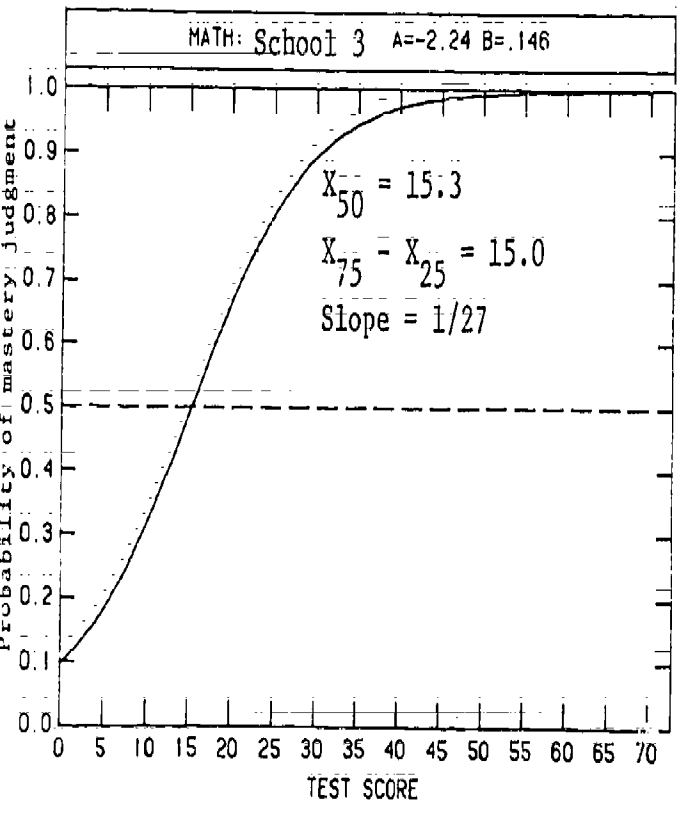


Figure 6c. Estimated relationship between math test scores and probability of mastery judgment: Schools 1-4.







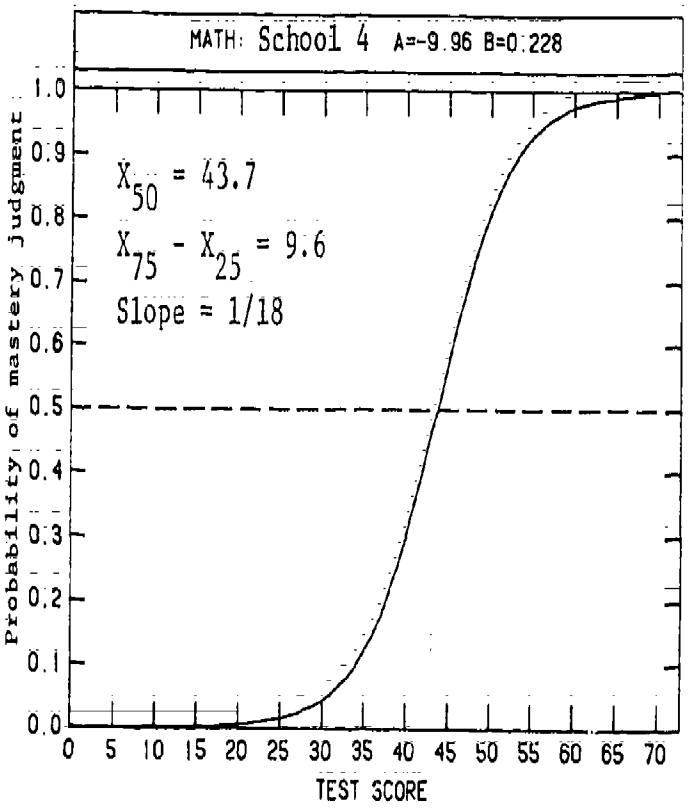
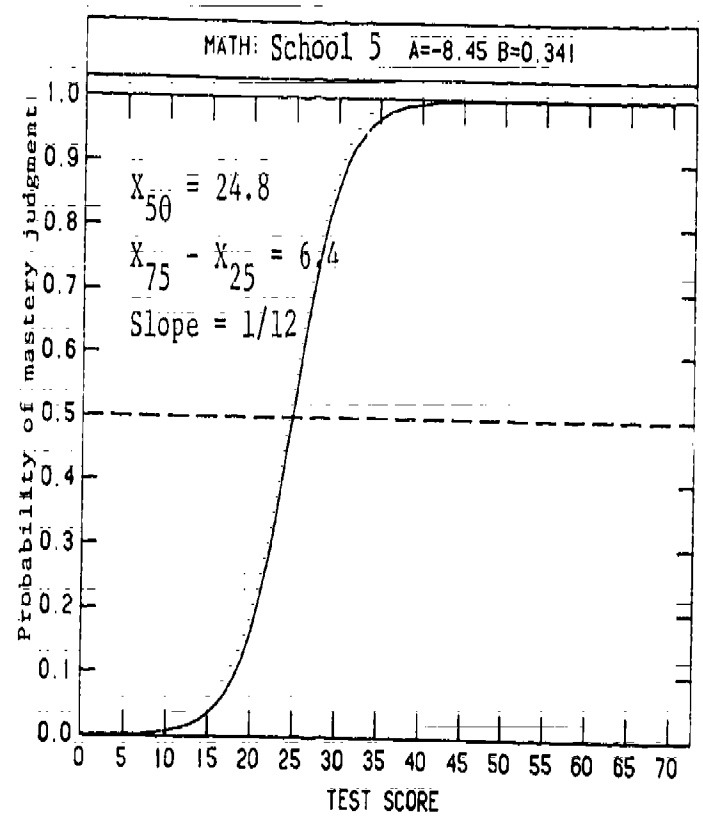
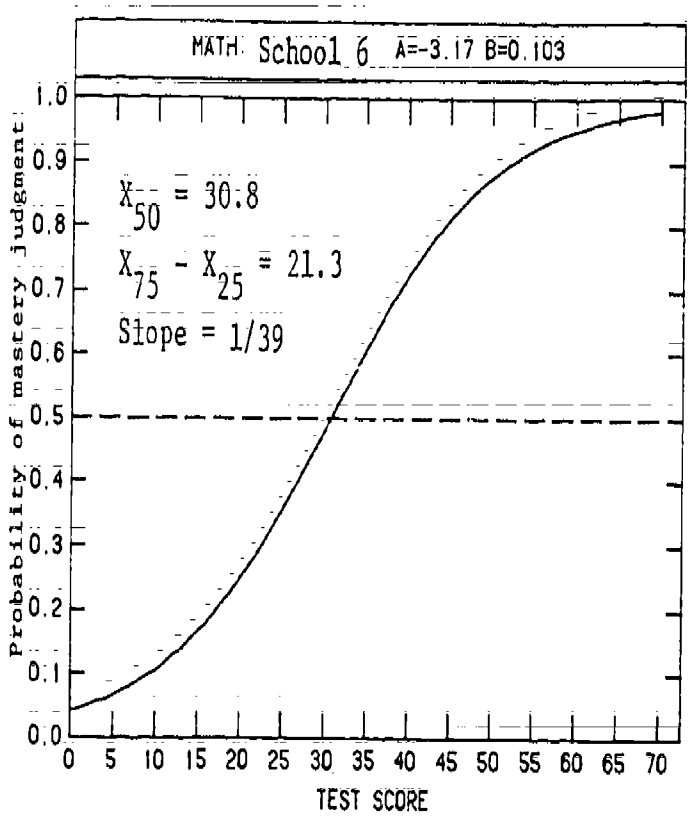
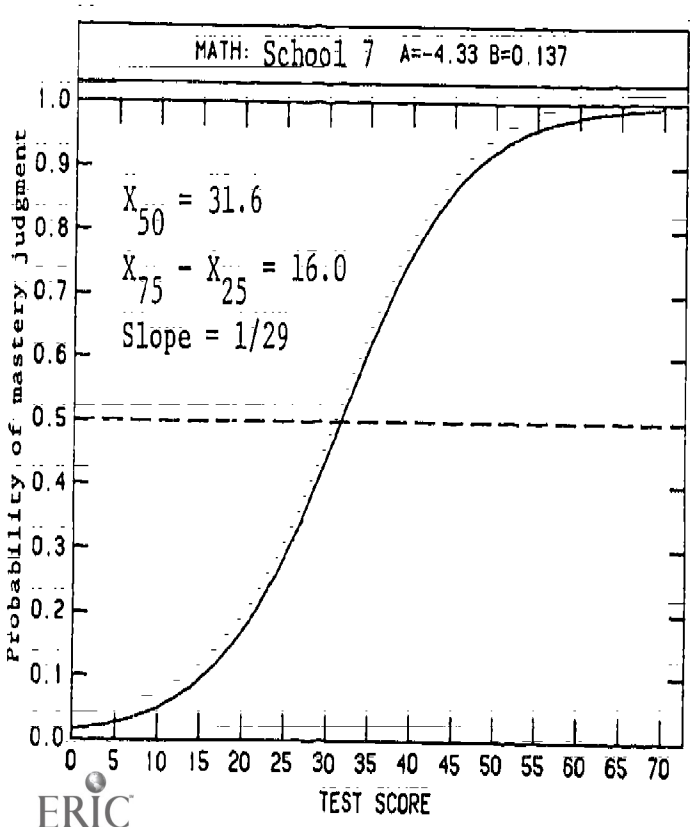


Figure 6d. Estimated relationship between math test scores and probability of mastery judgment: Schools 5-8.







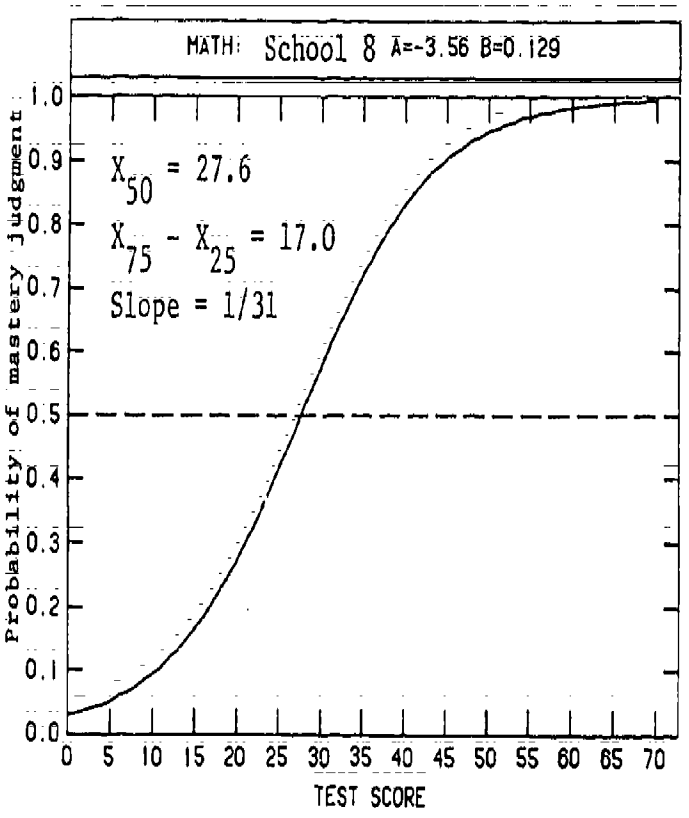
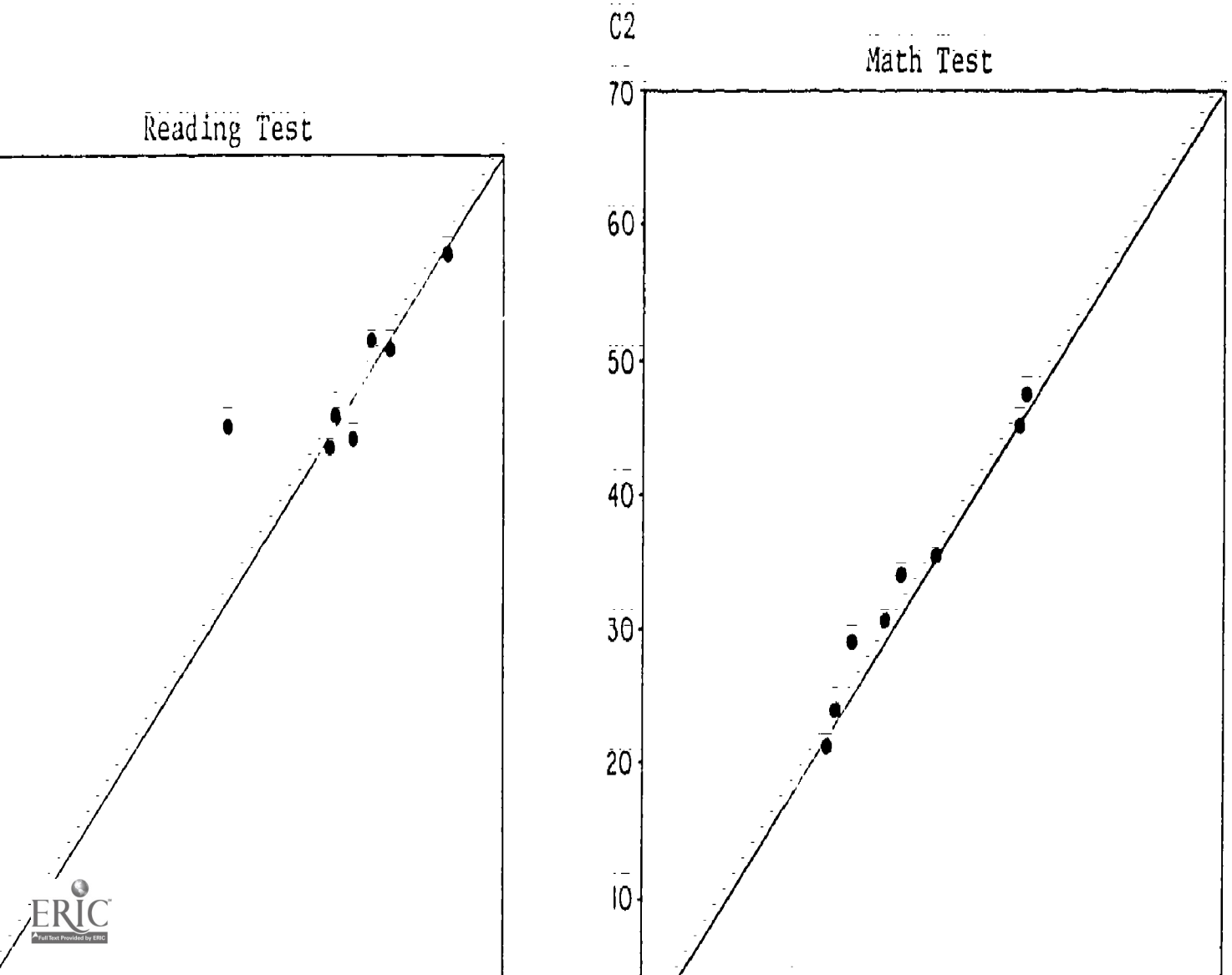


Figure 7 Borderline-group passing score and "C2" standard for each school.



Appendix

Computing Passing Scores from Contrasting-Groups Data.

The procedure used in this study to compute passing scores from contrasting-groups data is based on decision theory. This approach requires, for each test score level, an estimate of the probability that a student with that test score would be judged a master (given that the student would be judged either a master or a nonmaster). If the number of students at each score level were very large, it would make sense to use the percentage of masters at each score level as a probability estimate. But, because the number of students at each score level is small, it is necessary to use some other means of estimating the conditional probability function. The method used in this study is called "logistic regression". It assumes that the conditional probability function can be described by the equation

$$P = \frac{1}{1 + e^{-(a + bx)}}$$

where e is the familiar constant 2.71828, x is the student's test score, and a and b are parameters estimated from the data (in this case, by using the BMDP statistical software package). Once the a and b parameters have been estimated, it is possible to find the value of x (the test score) corresponding to any desired value of P (probability of being judged a master). In particular, when P = .50, then x = -a/b. We have referred to this score as "X_{50}". The slope of the conditional probability curve is steepest at this point, where it is equal to b/4.

A-2

Stability of the Contrasting-Groups Passing Score.

The sampling variability of X_{50} depends on the slope and the sample size. The larger the sample and the steeper the slope, the more precisely X_{50} can be determined. When the sample is not too small, it is possible to estimate the standard error of X_{50} by the formula

$$\text{Var}\ (X_{50}) = \frac{1}{b^2}\ \text{Var}\ (a) + \frac{a^2}{b^4}\ \text{Var}\ (b) + \frac{2a}{b^3}\ \text{Cov}\ (a,b).$$

The variances and covariance of the a and b parameters are computed by BMDP.

The standard errors of the contrasting-groups passing scores computed for each school as a whole were as follows:

		Reading (65 points)	Math (70 points)
School	1	1.0	2.6
	2	3.2	1.2
	3	not computed	2.6
	4	1.7	1.1
	5	6.2	0.7
	6	1.0	1.2
	7	3.7	1.2
	8	3.7	1.6

These standard errors do not include the selection of individual teachers as a source of variability. They do include any unreliability in the individual teachers' judgments and in the students' test scores, as well as the selection of individual students. Thus, the standard errors refer to replications of the procedure with the same teachers but different students.

A-3

The standard errors of the contrasting-groups passing score computed for individual teachers varied from 1.6 to 7.9 points for the reading test and from 1.4 to 5.1 points for the math test.

Stability of the Borderline-Group Passing Score.

Although the borderline group often contained students with very high and very low scores, its median could be quite stably estimated when the method was applied to the school as a whole. There is no single simple formula for the standard error of the median; its standard error depends on the parent distribution, which is unknown. However, the standard error of the mean should provide a reasonable approximation. The standard errors of the mean of the scores of the borderline group in each school were as follows:

		Reading (65 points)	Math (70 points)
School	1	0.9	1.0
	2	1.2	1.1
	3	not computed	0.8
	4	0.8	1.5
	5	1.4	0.7
	6	0.9	0.9
	7	1.4	1.8
	8	1.9	1.4

These standard errors refer to replications of the procedure with the same teachers but different students.

Table A1: Passing scores and observed mean and standard deviation of students' scores in each school.

		Passing Score				Mean Score	Standard Deviation
		Contrasting-Groups	Borderline-Group	Nedelsky	Angoff		
Reading							
School	1	43.9	47	38.1	--	45.5	11.1
	2	37.1	51	36.5	--	54.0	8.3
	3	a	a	46.0	--	39.1	8.9
	4	50.4	58	46.7[b]	--	58.4	7.1
	5	33.4	32	--	25.9	39.8	11.8
	6	44.2	45	--	32.9	44.1	12.9
	7	24.0	44	--	43.2	47.0	10.5
	8	34.4	48.5	--	37.6	53.8	8.5
Math							
School	1	41.4	29	25.0	--	28.6	9.4
	2	52.3	45	27.7	--	39.0	12.5
	3	15.3	22	29.8	--	22.7	6.6
	4	43.7	46	42.2	--	50.9	10.9
	5	24.8	23	--	24.6	24.7	8.9
	6	30.8	25	--	34.7[c]	28.3	9.7
	7	31.6	31	--	33.9	34.9	10.6
	8	27.6	35.5	--	48.6	39.4	11.9

[a] Could not be computed.

[b] Does not include one teacher who was unable to attend standard setting session.

[c] Does not include two teachers who were unable to attend standard setting session.

Table A2a. Means and Standard Deviations of Reading Test Scores, by School and Judgment

School		All Students	"Master"	"Borderline"	"Nonmaster"
1	Mean	45.5	52.8	46.2	36.5
	SD	11.1	9.3	9.7	7.6
	N	329	115	112	98
2	Mean	54.0	56.4	48.8	43.4
	SD	8.3	7.1	6.5	9.1
	N	186	137	31	16
3	Mean	39.1	39.8	--	38.5
	SD	8.9	8.5	--	9.2
	N	146	66	0	80
4	Mean	58.4	60.6	57.3	49.3
	SD	7.1	4.1	5.5	12.4
	N	204	129	49	25
5	Mean	39.8	47.9	32.5	42.0
	SD	11.8	8.6	9.1	9.9
	N	124	47	44	21
6	Mean	44.1	53.7	42.8	33.8
	SD	12.9	8.0	11.3	11.3
	N	387	130	152	102
7	Mean	47.0	49.4	40.8	37.0
	SD	10.5	9.6	7.6	11.2
	N	288	218	30	29
8	Mean	53.8	55.4	46.6	47.3
	SD	8.5	7.6	7.3	8.8
	N	284	226	14	41

Table A2b. Means and Standard Deviations of Math Test Scores, by School and Judgment

School		All Students	"Master"	"Borderline"	"Nonmaster"
1	Mean	28.6	36.5	29.8	25.9
	SD	9.4	8.1	8.4	8.9
	N	229	36	65	128
2	Mean	39.0	53.7	44.5	33.5
	SD	12.5	7.4	10.7	10.3
	N	376	51	95	230
3	Mean	22.7	24.1	22.3	20.0
	SD	6.6	7.8	5.6	4.0
	N	157	73	54	29
4	Mean	50.9	56.7	45.5	39.7
	SD	10.9	8.3	8.9	7.0
	N	209	127	34	46
5	Mean	24.7	32.6	24.0	18.9
	SD	8.9	8.7	6.9	5.3
	N	266	74	106	78
6	Mean	28.3	33.5	25.9	25.2
	SD	9.7	9.4	8.4	8.8
	N	413	145	94	173
7	Mean	34.9	40.0	32.7	29.2
	SD	10.6	9.7	10.4	7.8
	N	266	130	35	92
8	Mean	39.4	43.6	36.0	30.0
	SD	11.9	11.4	10.2	8.3
	N	275	166	52	56

References

Angoff, W. H. Scales, norms, and grade equivalent scores. In R. L. Thorndike (Ed.). Educational Measurement (2nd Edition). Washington, DC: American Council on Education, 1971, pp. 514-515.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement. 1980, 17, pp. 176-178.

Livingston, S. A. and Zieky, M. J. Passing Scores. Princeton, NJ: Educational Testing Service, 1982.

Mills, C. N. and Barr, J. E. A comparison of standard setting methods: Do the same judges establish the same standards with different methods? Paper presented at the annual meeting of the American Educational Research Association, Montreal, 1983.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, pp. 3-19.

Poggio, J. P., Glasnapp, D. R. and Eros, D. S. An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1981.

Shepard, L. Standard setting issues and methods. Applied Psychological Measurement, 1980, 4, 447-467.