DOCUMENT RESUME

ED 228 205 SP 022 052

TITLE Public School Teachers' Perceptions of Test Validity and Methods of Interpreting Test Score Data. Final Report.
INSTITUTION Mississippi State Univ., Mississippi State, Bureau of Educational Research.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Dec 82
GRANT NIE-G-81-0106
NOTE 139p.
PUB TYPE Reports - Research/Technical (143) -- Statistical Data (110)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Academic Standards; Attitude Change; Decision Making; Educational Research; *Educational Testing; Elementary Secondary Education; Predictive Validity; Public School Teachers; Student Evaluation; *Teacher Attitudes; *Testing Problems; *Test Interpretation; *Test Reliability; *Test Validity

ABSTRACT

Survey results have suggested that, while teachers like to have test information available, most do not have great skill or consistency in interpreting test score data. Teachers who consider a certain type of test very valuable or useful are less likely to question the accuracy of the scores than are teachers who consider a test to be of little value. Also, teacher judgments of test score accuracy are apparently tempered on the basis of other information. The overall intent of the project described in this report was to learn more about how teachers perceive test score data; how they use test and nontest data in decision making; factors which might be related to knowledge or perception of test score data; and whether knowledge or perception of test score data can be changed through brief instruction. Reported are six studies that were conducted, focusing on: (1) teachers' perceptions of test score accuracy; (2) teachers' interpretations of pupil performance record; (3) teachers' interpretations of point and interval score estimates; (4) teachers' estimates of costs and losses in decisions; (5) factors influencing teacher estimation of pupil performance; and (6) changes in teachers' knowledge and perceptions of test score data. A description is given of the methodology used, the results, and an interpretation of findings for each of the tests. Appendixes include instruments used in the studies and tables of results. (JD)

Public School Teachers' Perceptions of Test Validity and Methods of Interpreting Test Score Data

Final Report

December 1982

NIE-G-81-0106

David T. Morse
Project Director

Mississippi State University
Bureau of Educational Research and Evaluation

The activity which is the subject of this report was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

Overview of the Project

Background

There are today two major factors which have had, and will continue to have an enormous effect upon the shape and stance of education in America. These factors are accountability and evaluation, both instructional and program. One clear outcome of these factors' impact is an increase in educators' dependence upon tests to provide empirical data for making instructional or programmatic decisions or changes. It is fairly easy to see the rapid increase in state-sponsored every-pupil testing programs as but one example. Shoemaker (1978) indicated that the number of states endorsing and mandating some form of state-wide academic skills assessment has risen from thirty to now over forty since the 1976-77 academic years. That the public is concerned with the measured capabilities of students in all or some of the traditional school content areas is underlined by the considerable volume of literature devoted to competency testing (c.f. Phi Delta Kappan, May, 1978).

Even if the tests to be used meet stringent standards, such as those used by the Center for the Study of Evaluation (Hoepfner et al., 1972), or by any reviewer in the Mental Measurements Yearbook (Buros, 1978), for norm-referenced tests, or those proposed by Popham (1978) or by Hambleton and Eignor (1978) for criterion-referenced tests, there is no guarantee that the users will be able to interpret the results of the tests sensibly. That is, a school or school district could conduct a study which met all the criteria set forth by the joint dissemination review panel publication, the JDRP Ideabook (Tallmadge, 1977), and yet be a failure in terms of dissemination if the results released to local teachers and administrators were mis-interpreted. The simple fact that most commercial test publishers will provide, upon request, grade-equivalent scores -- in spite of their many technical flaws and susceptibility to mis-interpretation (Hills, 1981; Tallmadge & Horst, 1974) -- should alert the reader that the state of the art in test score interpretation perhaps lags too far behind the level necessary for sound decision-making. Two recent surveys, conducted independently, canvassed the local coordinators of accountability or testing in each school district in two states. One question on both surveys asked these district coordinators to estimate what percentage of teachers in their district could interpret a grade-equivalent score properly. For Florida, the median estimate was 50% (Hills, 1977), while for Mississippi, the median estimate was 40% (Morse, 1978). Further, when these coordinators were asked to cite instances of the worst mistakes in teachers' use of tests and measurements, examples of mis-interpretation of test scores were given most often in Hills' survey (1977) and nearly most often in Morse's survey (1978).

Large-scale surveys of teachers have been conducted over the years and, as a set, suggest that teachers like to have test information available but don't have great skill or consistency in interpreting test score data (Hastings et al., 1960; Goslin, 1967; Rudman et al., 1980; Burry et al., 1982; Kellaghan, Madaus & Airasian, 1982).

Some Results of Interpretation Studies

Fleming and Antonen (1971), in a study designed to replicate the findings of Rosenthal and Jacobson (1968), were not able to induce the type of expectancy effects which the "Pygmalion" study purported to obtain. However, the study did show significant differences in the degree of accuracy (or validity) which public school teachers were willing to ascribe to sets of supplied test scores (either IQ, ability percentile, or an inflated IQ for their pupils), based on the degree of utility or value they attributed to the particular test type. That is, teachers who considered a certain type of test very valuable or useful were less likely to question the accuracy of the scores than were teachers who considered the particular test to be of little value.

There is evidence that giving test scores in the form of confidence bands reduces the amount of precision teachers will ascribe to a test. Beggs, Mayer, and Lewis (1972) gave teachers either: (a) an IQ score; (b) an IQ score with an explanation of test reliability and validity; (c) an IQ score with a prediction of future work; or (d) a percentile band (confidence band) for the IQ score, with an explanation of test reliability and validity. Over a four-month period, teachers' estimates of the accuracy of the scores was much more consistent for those given IQ scores (conditions a,b) than for those given a confidence band (condition d). Also, when the teachers were asked to estimate their pupils' actual, as opposed to measured, IQs, the teachers given the confidence bands were again less consistent over the same period than were those given the IQ scores (conditions a,b). Morse (1964) gave undergraduate students hypothetical test scores expressed either as a percentile rank, narrow percentile band (± .5 SD), or wide percentile (± 1.0 SD). In nearly all cases, the respondents perceived the percentile rank as being significantly further from the mean (of 50) than was its corresponding narrow percentile level, which was in turn perceived as being significantly further from the mean than its corresponding wide percentile band. In other words, the more realistically the accuracy of the hypothetical test was represented (by increasing the confidence band), the less willing were the respondents to suggest that the given scores differed from the mean. Thus, for genuinely low scores, the respondents were in fact over-estimating the relative position, while for genuinely high scores the respondents would under-estimate the relative position.

Teachers apparently temper judgments of test score accuracy on the basis of other information. Frederickson and Marchie (1966) gave a small group of teachers hypothetical protocol data including an IQ score, aptitude score, and basic skills score of a pupil's class performance. Teachers were asked whether they should accept the score as valid, question its validity, or make no judgment. The highest acceptance rates were noted for high scores versus average class performance, while the lowest acceptance of the test scores as being valid was noted for average or low IQ score versus high class performance. While this indicates some questioning of test score accuracy under certain conditions, it also indicates that the tests are being perceived as valid predictors of class performance. This explains why teachers might feel comfortable knowing that a student whose class performance is average may have a high IQ score (e.g., an "underachiever"), and feel less comfortable when told that a student whose class performance is high may have an average or low IQ (e.g., implying that teachers do not accept the notion of an "overachiever"). Leither (1976) found when teachers were given achievement test scores of their students in percentile rank units, it was not uncommon for them to draw upon their knowledge of unrelated student background information in order to interpret the scores. This was the outcome in spite of the fact that the teachers were asked merely to interpret the pupils' performance on the test.

Practical Significance of the Problem

The Title I evaluation models, now required for use in all ESEA Title I program evaluation, call for the use of a placement or selection test, an evaluation instrument, and have introduced an entirely new score metric, the normal curve equivalent (NCE). A large number of decisions about individual students, based on test scores, must surely be taking place almost daily as a result. While widely critiqued for its lack of experimental rigor since publication, the manuscript Pygmalion in the classroom (Rosenthal & Jacobson, 1968) raised some interesting questions as to how teachers use test score information, whether consciously or not. If many teachers are not able to interpret test scores properly, as suggested above, and if there is but one grain of truth to the "Pygmalion" notion that test scores are accorded an inordinate amount of weight by teachers, the need for a study of how teachers interpret different test score data, and whether this capability may be improved through brief training ought to be very clear. In all likelihood, compensatory programs such as Title I result in test scores being used far more often in making selection and placement decisions. If the decision-makers have not learned how to interpret test score information properly, then many students stand to suffer.

Project Description

This project was composed of six separate substudies, each of which was initiated to answer one or more specific research questions. Overall, a total of 474 public school teachers from twenty-six school districts in Mississippi participated in one or more phases of the study. The purpose and focus of each study is explained below. The overall intent was to learn more about how teachers perceive test score data; how teachers use test and nontest data in decision-making; factors which might be related to knowledge or perception of test score data; and whether knowledge or perception of test score data can be changed through brief instruction. While possibly raising more questions than are answered, the substudies do appear to indicate likely areas for future research.

I. Teachers' Perceptions of Test Score Accuracy

The purpose of study was to investigate how teachers choose to use and interpret test information. The study included an examination of: (a) How perceived validity of test scores is affected by the congruence of test and nontest information; (b) The relative perceptions of test score scale accuracies; and (c) The relative perceptions of the utility of various types of test and nontest information for making placement decisions. The object of the study was to allow the examination of what types of data teachers choose to use as well as how test and nontest performance information are considered and combined.

II. Teachers' Interpretations of a Pupil Performance Record

The purpose of this study was to investigate how teachers interpret pupil performance record data. The study included an examination of: (a) Which of the available types of performance measures available teachers use in drawing initial judgments of pupil performance; and (b) The type of performance measure teachers believe to be the most reliable. The object of the study was to allow the determination of what types of data serve to mediate judgment of a student's capability and what type of data is thought to be most trustworthy.

III. Teachers' Interpretations of Point and Interval Score Estimates

The purpose of this study was to investigate how teachers perceive test scores depending upon whether a point or interval estimate is provided. Specifically, the study included an examination of: (a) Whether practicing educators interpret point and interval estimates differently; (b) Whether the width of an interval estimate affects the resulting perception of a score; and (c) Whether any systematic trends in the types of scores could be discerned. The object of the study was to allow the determination of whether reporting point or interval estimates of performance would result in different perceptions or interpretations of the scores.

IV. Teachers' Estimates of Costs and Losses in Decisions

The purpose of this study was to investigate how teachers perceive the costs or losses associated with incorrect decisions or outcomes, and how these relate to the judged likelihood of such outcomes. The study included an examination of: (a) The frequency with which standardized achievement test scores accurately estimate pupil skill, as judged by teachers; (b) The types of outcomes or decisions which teachers perceive to be less desirable; and (c) The relative rates of incorrect decisions or outcomes being made, as judged by teachers. The object of the study was to allow the determination of what type of decision-making system might be judged to operate in education settings.

V. Factors Influencing Teacher Estimation of Pupil Performance

The purpose of this study was to investigate some of the factors governing the dynamics of how teachers interpret performance information in making decisions about pupils. These factors included: (a) valence of information (positive or negative); (b) congruence of follow-up information with initial performance; (c) reliability of information; and (d) gender of the pupils. Information protocols for hypothetical students were presented to teachers, and teachers were asked to judge, on the basis of information given, the chances of the "student" succeeding in school. The object of the study was to allow the examination of how much impact, if any, the four factors may have in the types of judgments teachers make concerning students, as well as whether selected teacher characteristics make any difference in the observed judgments.

VI. Changes in Teachers' Knowledge and Perceptions of Test Score Data

The purpose of this study was to investigate whether, and to what degree, measured knowledge or perceptions of test score data can be changed as a result of short-term, directed training. The study was designed to allow an examination of: (a) What changes in knowledge or perception of test score data could result from short-term training; and (b) Whether differences could be detected which were attributable to teacher characteristics of measurement background, certification or teaching level. The object of the study was to determine the efficacy of a modest training intervention in measured knowledge or perceptions teachers possess concerning test score data.

Substudy I

Teachers' Perceptions of Test Score Accuracy

How teachers choose to use and interpret test information is an aspect of educational practice which has not been extensively researched. Prior research results suggest that the degree to which test data are used by teachers depends upon how accurate or dependable the scores or data are perceived to be. This study was initiated to investigate three specific aspects of how teachers perceive test score data: (a) How do perceptions of test data accuracy vary as a function of the congruence of test data with other performance indicators? (b) Which types of common score scales do teachers believe most accurately summarize test performance? (c) Of the various types of test and nontest data which may be used for making placement decisions, which do teachers believe to be most accurate? These three research questions serve to define the scope of the present study. Given the present level of understanding of how decisions are made, the answers to these questions could provide insight as to how and what kind of test data should be presented to enhance the likelihood of sound use.

Methodology

Sample

Participants were 143 public school teachers from fourteen different school districts in Mississippi. These participants were in attendance at a workshop on test development. About 82% were female and 18% were male. The school districts represented were from the western, central and northeastern portions of the state.

Instruments

Data for the first research question came from participants' responses to a set of items asking the reader to judge the validity of a given test score, in light of other known, nontest information. Each respondent was presented eight such items (there were sixteen different items in all). The items presented various combinations of test score and nontest score data. Nontest score data were such data as marks in a given course. In each item, respondents were asked to judge the test score as valid, questionable or invalid. Items were classified as congruent if both the test and nontest data were high or low. However, incongruent combinations (e.g., high test score presented with low nontest score) were also included.

An example of a congruent (high test; high nontest score) item is: A female student, eighth grade, has an average grade of A. Her

new CAT-77 reading comprehension percentile rank is 91. This score is:
a. Valid
b. Questionable
c. Invalid

An example of an incongruent (low test score; high nontest score) item is:
A female student, twelfth grade, has a semester average of 93 in Senior English. Her new CAT-77 language arts percentile rank is 30. This score is:
a. Valid
b. Questionable
c. Invalid

Internal consistency reliability for this measure, estimated by coefficient alpha, was .80. A copy of the full set of items is presented in Appendix A.

Data for the second and third research questions came from separate pair-comparison questionnaires. The first presented five types of test score scales, including: Raw score (number right); Percentile rank; Grade-equivalent score; CAT-77 ADSS (a proprietary scale score); and Stanines. These five score types represent perhaps the most widely used -- exclusive of the NCE scores -- score scales in Mississippi. The second questionnaire presented seven sources of information which could possibly be used in making pupil placement decisions. These included: Prior course grades or marks; Standardized achievement test scores; Prior teacher's written recommendation; Individual I.Q. test; Prior school counselor's written recommendation; Local criterion-referenced (CR) achievement test scores; and Parents' description of child's school accomplishments. A copy of these instruments is included in Appendices B and C, respectively.

The method of pair comparisons requires that all possible pairs of stimuli be presented in a forced-choice format; the respondent must select one as preferable to the other. This method permits, if the necessary assumptions hold, interval scaling of the relative positions of the stimuli (Guilford, 1954). The order, sequence and pairing of stimuli were generated by use of a random number table, the intent being to avoid possible position bias.

For each questionnaire, respondents were told that there were no "right" or "wrong" answers and that they should respond on the basis of their own beliefs.

Specific instructions for the test score type questionnaire were:

Each year, the state sponsors testing of students in grades 4, 6 and 8 in basic skills on the California Achievement test. Various

> types of scores are provided for students who take the test. For each of the following items, please select the type of score you believe would best help you, as an educator, to make sound decisions about what a student had or had not learned.
>
> Please circle the letter of the type of score you select for each item.
>
> Remember, you should choose the type of score YOU think would best help in making sound decisions about a student's skills.

Specific instructions for the types of data questionnaire were:

> When a new student comes to your school, some type of placement decision must be made. For each of the following questions, please circle the letter of the type of information you believe is likely to be MOST ACCURATE for making sound placement decisions.

All respondents were able to complete the longest questionnaire easily within fifteen minutes. Only one questionnaire was administered a day.

Results

Question 1: How do perceptions of test data accuracy vary as a function of the congruence of the test data with other performance indicators?

Summary statistics by possible congruence category are presented in Table I-1. Higher scores represent greater perceived validity for the category of interest. Scoring was on a simple three-point scale, "valid" was assigned three points, a "questionable" rating was given two, and "invalid" was scored as one point. From the results in Table I-1, the reader may deduce that test information which was congruent (e.g., low-low or high-high) was perceived as more valid than was the test score information which was incongruent. There was a sizable advantage in ratings for congruent and high score data over those for congruent and low score data. For the incongruent data, there was a slightly greater tendency for the respondents to consider high test-low nontest matches as more believable than low test and high nontest combinations. The magnitudes of these differences are presented in Table I-2. The effect sizes shown in Table 2 range from small (.27) to very large (1.53). The overall hypothesis of equal ratings among the congruence categories was rejected at traditional alpha levels (F=119.51; df=3/1097; p<.001).

Question 2: Which types of common score scales do teachers believe most accurately summarize test performance?

TABLE I-1

Summary of Congruence Category Means

		Nontest Data	
		Low Score	High Score
Test Data	Low Score	2.27	1.81
		(0.74)	(0.67)
	High Score	1.98	2.73
		(0.59)	(0.52)

NOTE: Figures in parentheses are standard deviations; all values based on 143 cases.

TABLE I-2

Summary of Effect Size Estimates for Congruence Categories[1]

	Category[2]		
Category	Low-High	High-Low	High-High
Low-Low	.65	.43	.72
Low-High	--	.27	1.53
High-Low		--	1.35

[1] Effect size defined as $(\overline{X}_1-\overline{X}_2)/S_{pooled}$; all values based on 143 cases.

[2] Categories represent specific test-nontest score combinations.

Table I-3 includes the results of the pair comparison judgments. Overall, grade-equivalent scores were judged to be most accurate by the teachers, followed by percentile ranks. Further behind, and nearly equal in ranking, were raw scores and proprietary scale scores. Bringing up the distant rear was stanines. The scale values may be interpreted in a relative sense; that is, grade equivalent scores were preferred about twice as much as raw score and scale scores, and were about six times more popular than stanines. Shifting to a different scale (T-scores) removes the "anchor," but still permits relative contrasts. It is interesting to note that one of the least sound score scales is considered by teachers as most useful for making decisions about pupils. On the other hand, the stanine, which was designed to reflect the inherent uncertainty in a point estimate of an examinee's score, is least preferred.

Question 3: Which types of data do teachers believe to be most accurate for making placement decisions?

The pair comparison judgment results are summarized in Table I-4. Overall, test scores based on an achievement measure (both "standardized" and "local CR") are given highest ratings. After these comes student grades, then written recommendations by teacher and school counselor, respectively. Individual IQ test results ranked below the previous five. Finally, considerably below IQ scores was the parents' recommendation. Perhaps teachers have had much experience with parents' judgments of their child's capacity, and have found it wanting.

That performance-based measures should be accorded high ranks seems reasonable, given that prior performance -- such as grade point average -- is typically the best single predictor of future performance. What is intriguing is the fact that IQ tests, though a specialized performance measure, are possibly perceived as not sufficiently relevant to use in placement decisions, if other alternatives exist.

Summary

Test data are apparently more readily accepted if: (a) congruent with known nontest data; or (b) high rather than low if incongruent. That is, the so-called "under-achiever" (one who performs below the level at which a test might indicate is possible) is perhaps slightly more acceptable than is the notion of an "over-achiever." If given their choice, the participants in this study would much rather have grade-equivalent scores provided for their use than most others -- this in spite of the fact that possibly few people could give an accurate paraphrase of how one may interpret a grade-equivalent score. Finally, performance data are perceived as preferable to nonperformance data for making sound placement decisions.

TABLE I-3

Summary of Scale Values for Test Score Types

Test Score Type		Scale Value	T-score
Grade-equivalent Score		6.15	61
Percentile Rank		5.41	57
Raw Score (number right)		3.84	49
Scale Score (CAT-ADSS)		3.41	47
Stanine		1.00	35
	Mean	3.96	50
	S.D.	2.00	10

NOTE: All values based on 143 cases.

TABLE I-4

Summary of Scale Values for Data Sources

Data Source	Scale Value	T-score
Standardized achievement test scores	6.64	57
Local CR achievement test scores	6.59	57
Grades or marks	6.43	56
Teacher's written recommendation	5.85	53
Counselor's written recommendation	5.45	51
Individual IQ test	4.61	47
Parents' description of child's accomplishments	1.00	29
Mean	5.22	50
S.D.	2.00	10

NOTE: All values based on 143 cases.

These results suggest that teachers are not only willing to make use of test data, but might actually prefer it to other data types. However, if some of the outcomes observed in this study carry over to the classroom, the reader may well wonder whether test data are being used in a sound fashion. The challenge to both researchers and publishers should be clear: To develop a useful and sound means for teachers to move from pupil results to considered decisions which will best facilitate each child's educational success.

Substudy II

Teachers' Interpretations of a Pupil Performance Record

The results of substudy I suggest that certain score scale types are preferred by teachers to others, as are certain sources of pupil performance information. Also, the perceived validity of test scores will vary as a function of the congruence of test and nontest score data, as was found in another study by Frederickson and Marchie (1966). Farr and Griffin (1973) and more recently, Newman and Stallings (1982) suggest that teachers' awareness of sound measurement practice may have implications for their classroom assessment and decision-making behavior. This study was initiated to answer two specific aspects of how teachers interpret pupil performance record data: (a) What type(s) of performance indicator do teachers use in drawing initial judgments of pupil performance? (b) What type of performance indicator do teachers believe to be the most trustworthy? These two research questions serve to define the scope of the present study. The results of this brief, exploratory study were used to shape the work described in substudies V and VI.

Methodology

Sample

Participants were 210 public school teachers from sixteen different school districts in Mississippi. These participants were in attendance at a workshop on interpretation of test score data, which was a part of a week-long workshop on test development. About 76% were female and 24% were male. The school districts represented were from the western, central, and southern portions of the state.

Instruments

Data for both questions came from participants' responses to two items which followed a hypothetical pupil performance record. There were two versions of the protocol used, varying primarily in terms of what IQ score was affixed to the record. Other performance data included semester grade averages and standardized achievement test scores, expressed in percentiles and scale scores. (NCE scores were also included on the first record.) The hypothetical pupil records are included in Appendix D.

For each record, respondents were told that there were no "right" or "wrong" answers and that they should respond on the basis of their own beliefs.

Specific instructions were:

The following information has come from an anonymous student's cumulative record. Please examine it carefully and answer the questions which follow.

All respondents were able to complete the task easily within ten minutes.

Results

Question 1: What type(s) of performance indicator do teachers use in drawing initial judgments of pupil performance?

Answers to item 1 were coded so that an answer of "Well above her ability" was coded as a 3, "About equal to her ability" as a 2 and "Well below her ability" was coded as a 1.

Differences on the first item responses between protocol groups are summarized in Table II-1. The effect size of the difference in ratings was 0.75 standard deviations (based on pooled variance estimate), which was statistically significant at the .05 level ($F_{1,208} = 35.31$). Because the pupil performance protocols differed primarily on the stated IQ score, a reasonable conclusion is that one of the least favored score types, IQ, is given most weight in judging performance relative to "ability." A second possible interpretation is that the respondents paid close attention to the directions and concluded, correctly, that IQ data was the measure most indicative of ability. However, in most tests and measurements courses, the concept of errors of measurement is presented; so-called "normal" ranges for IQ are generally described as between 90-110. The results suggest that these two pupil records are not at all perceived as equivalent in ability.

TABLE II-1

Summary Statistics for Protocol Groups on Performance Judgment

	Group 1 (Low IQ Protocol)	Group 2 (High IQ Protocol)
Mean	2.23	1.82
Standard deviation	0.42	0.48
n	66	144

Question 2: What type of performance indicators do teachers believe to be the most trustworthy?

A summary of the options presented in item 2 suggests that teachers believe, and very likely are correct, that grade point average is the most reliable of the performance indicators listed on the protocol (50%). The next most popular choice was that of achievement test scores (31%). The IQ scores were about even in their rate of selection, about ten percent each (19% total). A chi-square test of independence for performance score choice and assigned protocol yielded no significant relationship (chi-square corrected = 1.82; 3 df; probability = .615). Thus, regardless of this hypothetical student being considered a relative "underachiever" or "overachiever," teachers did not vary their perception of grade average as the most reliable of the given performance data.

Finally, similar contrasts for questions 1 and 2 were conducted for teacher gender, test course status (yes or no: Have you ever taken a course in tests and measurements?), certification (A, which is a B.S. level; AA, which is the M.S. or M.Ed. level; and AAA, which represents the Ed.S. level of training) or level of teaching (elementary, secondary or both). In all cases, no statistically significant differences were detected.

Summary

A considerable number of the participants judged the hypothetical student as an "overachiever" or an "underachiever," depending upon which of two performance protocols was assigned. These judgments would appear to be based primarily upon the listed IQ score in relation to the other performance data. Yet, IQ was listed as the least reliable of the types of information available and, from substudy I, is one of the less-preferred data sources. Given that teachers can form opinions of pupil performance, the questions which remain to be answered are: (a) Do these pre-conceived judgments of a pupil transfer to decisions made about the pupil? (b) Would these judgments be made on the job, or only in contrived tasks such as the one used in the present study? (c) How often do teachers believe their judgment, however formed, might be incorrect? Finally, (d) How long would a teacher have to observe a pupil in order to alter an initial judgment of the child if that judgment was incorrect?

These questions, if investigated, could serve to support the formation of what might be considered essential training in sound placement and decision-making principles for educators.

Substudy III

Teachers' Interpretations of Point and Interval Score Estimates

Among the possible conclusions of substudy II was the idea that teachers may not understand or may ignore the concept that test data are subject to "wobble" and that scores which are different may not be significantly different. Most measurement texts (e.g., Hills, 1981; Anastasi, 1976) suggest that confidence bands represent more realistically the degree of precision with which a test can estimate an examinee's true skill level. As a test's reliability increases, resulting confidence bands for any given confidence level will decrease in their width. Thus, wide confidence bands should be a tip-off to relatively low test reliability; armed with this information, users should be wary of placing considerable stock in wide confidence bands.

Morse (1964) investigated the differences in how undergraduate students, early in a course on tests and measurements, interpreted point and interval estimates relative to the mean score. His findings suggest that interval estimates (confidence bands) were more likely to be judged as closer to the mean than were point estimates (individual percentile ranks). Further, the phenomenon was more pronounced for "wide" confidence bands (± one standard deviation) than for "narrow" (± one-half standard deviation).

The present study was initiated to answer three specific questions: (a) Do practicing educators interpret point and interval score estimates differently? (b) Does the width of an interval estimate result in different perceptions? (c) Are there identifiable trends in the interpretations of these scores? The answers to these questions would have implications for both reporting practice and possibly for pre-service or in-service training needs of educators.

Methodology

Sample

Participants were 105 public school teachers from Mississippi, representing eleven different school districts. Of these, approximately 78% were female and 22% were male. The participants were attending a workshop on interpretation of test scores, which was part of a larger workshop on test development. The school districts represented were from the western, central and northeastern regions of the state.

Instruments

A single instrument was used to gather the data for questions 1-3. This instrument consisted of four sets of nine scores; either percentile ranks or percentile bands, one set to a page. To the side of each score was a rating scale which ranged from 1 to 5 for which the following key was given:

5 = Score is well above mean.
4 = Score is somewhat above mean.
3 = Score is equal or nearly equal to mean.
2 = Score is somewhat below mean.
1 = Score is well below mean.

Overall directions for the task were as follows:

Directions

On the follow sheets, you will find a number of test scores, expressed as percentile ranks or percentile bands.

Your percentile rank tells the percentage of a norm group that you have equaled or surpassed. For example, if your percentile rank for height in this class is 75, then you are as tall or taller than 75% of the persons in the class.

Because test scores tend to vary somewhat due to such chance factors as a lucky guess or the choice of questions, we sometimes express a score as a percentile band. The percentile band 50-75, for example, would mean that we are reasonably confident that the person earning this score is really better than the lower half of the group, but not as good as the top quarter of the group.

When the signal is given, open your booklet to page 1, and begin to work. Be sure that you finish each page before going on to the next page. DO NOT TURN BACK TO A PAGE ONCE YOU HAVE LEFT IT. WAIT FOR THE SIGNAL TO START.

The scores selected represented values of -2.0, -1.5, -1.0, -0.5, 0.0, +0.5, +1.0, +1.5, and +2.0 standard deviations from the mean, expressed as percentile ranks or as 67% confidence bands for various reliabilities. The wide confidence band, defined as ±1.0 standard deviations, assumes a test with zero reliability (e.g., standard error of measurement = standard deviation). The narrow confidence band, defined as ±0.5 standard deviations, assumes a test with reliability of .75. The very narrow band, defined as ±0.33 standard deviations, assumes a test reliability of .89. If one assumes that teachers make decisions from standardized achievement tests, then the very narrow band might be a realistic interval estimate. For locally constructed tests, the narrow or a "mid-wide" band might be more realistic.

The order of the scores was randomly determined and kept constant for each set. The sequence of the sets was randomized for all participants so as to avoid any order effects from biasing the results.

The ratings were then summed selectively for each set. Ratings on the four scores or bands based on the scores above the mean were summed for an "above the mean" total for each set. Similarly, ratings on the four scores or bands below the mean were summed for a "below the mean" total for each set. The rating scale being from one to five, each summed value had a potential range of from five to twenty. High values would suggest a perception of the scores being above the mean, while low values would indicate a perception of the scores being below the mean.

Internal consistency reliability estimates for the instrument were: (a) for the individual scores, alpha = .85 (k = 36); (b) for the "below the mean" sums, alpha = .66 (k = 4); and (c) for the "above the mean" sums, alpha = .78 (k = 4). A copy of the complete instrument is presented in Appendix E.

All participants were able to complete the instrument easily within thirty minutes.

Results

Question 1: Do practicing educators interpret point and interval score estimates differently?

A repeated-measures analysis of variance (ANOVA) was calculated for the four below the mean sums (one from the percentile rank set, one from the 1/3 S.D. band set, one from the 1/2 S.D. band set and one from the 1 S.D. band set). The results are presented in Table III-1. A statistically significant between-sets F-ratio was obtained, which suggests that, for the below the mean scores, there was a difference in how near to or far from the mean point and interval estimates were perceived to be. A similar analysis was calculated for the four above the mean scores, and it also resulted in a statistically significant between-sets F-ratio. The summary of that ANOVA contrast is presented in Table III-2.

Summary statistics for the summed scores are presented in Table III-3. There is a systematic change within each score type. The below the mean score sums tend to increase as the interval estimate becomes wider. (A value of 12 would represent a rating of the scores as being equal to the mean.) The opposite is true for the above the mean scores. As the interval estimate becomes wider, the summed ratings declined.

TABLE III-1

Repeated Measures of ANOVA Contrasts of
Sets of Percentile Ranks below 50

Source of Variation	Sum of Squares	df	Mean Square	F	Probability
Between Persons	248.89	104	2.39		
Within Persons	351.25	315	1.12		
Between Sets	95.00	3	31.67	38.55	.000
Residual	256.25	312	0.82		
Total	600.12	419			

TABLE III-2

Repeated Measures of ANOVA Contrasts of
Sets of Percentile Ranks above 50

Source of Variation	Sum of Squares	df	Mean Square	F	Probability
Between Persons	478.01	104	4.60		
Within Persons	471.50	315	1.49		
Between Sets	161.57	3	53.86	54.22	.000
Residual	309.93	312	0.99		
Total	949.51	419			

TABLE III-3

Perceived Distances of Scores from 50th Percentile

	Score Set			
Scores	Percentile Rank	1/3 Standard Deviation Band	1/2 Standard Deviation Band	1 Standard Deviation Band
Values below 50	5.27 (0.78)	5.60 (1.00)	5.82 (1.08)	6.56 (1.45)
Values above 50	17.71 (1.13)	17.36 (1.24)	17.53 (1.21)	16.10 (1.81)

Note: Values in parentheses are standard deviations; all values based on 105 respondents.

Question 2: Does the width of an interval estimate result in different perceptions?

Orthogonal contrasts were calculated for each score type and indicate, for the below the mean sums that: (a) the interval estimate ratings were perceived as significantly closer to the mean than the point estimate ($F = 50.69$; $p < .001$); (b) there was no difference between the very narrow and narrow interval estimates ($F = 3.07$; $p = .078$); and (c) the narrow and very narrow interval estimates were perceived as further from the mean than was the wide interval estimate ($F = 61.95$; $p < .001$).

Very similar conclusions could be drawn for the above the mean score set contrasts: (a) the point estimate was perceived as being significantly further from the mean than were the interval estimates ($F = 39.39$; $p < .001$); (b) there was no significant difference in perception of the very narrow and narrow interval estimates ($F = 1.55$; $p = .211$); and (c) the wide interval estimates were judged to be significantly closer to the mean than were the narrow and very narrow estimates ($F = 121.77$; $p \leqslant .001$).

Thus, while the teachers in this study did apparently interpret point and interval estimates differently, they did not distinguish systematically between the very narrow and narrow confidence bands. The wide interval bands, though, were perceived as significantly closer to the mean than the other two interval estimates.

Question 3: Are there identifiable trends in the interpretations of these scores?

Orthogonal tests of trend were calculated using polynomial coefficients from Winer (1971). The results of these contrasts are presented for the below the mean scores in Table III-4 and for the above the mean scores in Table III-5.

For the below the mean scores, there was a significant linear trend and an arguable quadratic trend ($F = 5.36$; $p = .020$) beyond the linear trend, depending upon the reader's preferred level of significance. The cubic trend was not statistically significant. For the above the mean scores, the linear, quadratic and cubic trends were statistically significant. These trends are illustrated in Figures III-1 and III-2, respectively.

Figure III-1 is suggestive of a linear trend for the below the mean scores, in which ratings approach the mean as one moves from a point estimate to increasingly wider interval estimates. Figure III-2 is suggestive of a cubic trend, thanks mostly to a dramatic change for the wide band ratings. Again, as one changes from a point estimate to increasingly wider interval estimates, the assigned ratings decline

Figure III-1
Mean Ratings for Scores Below Mean

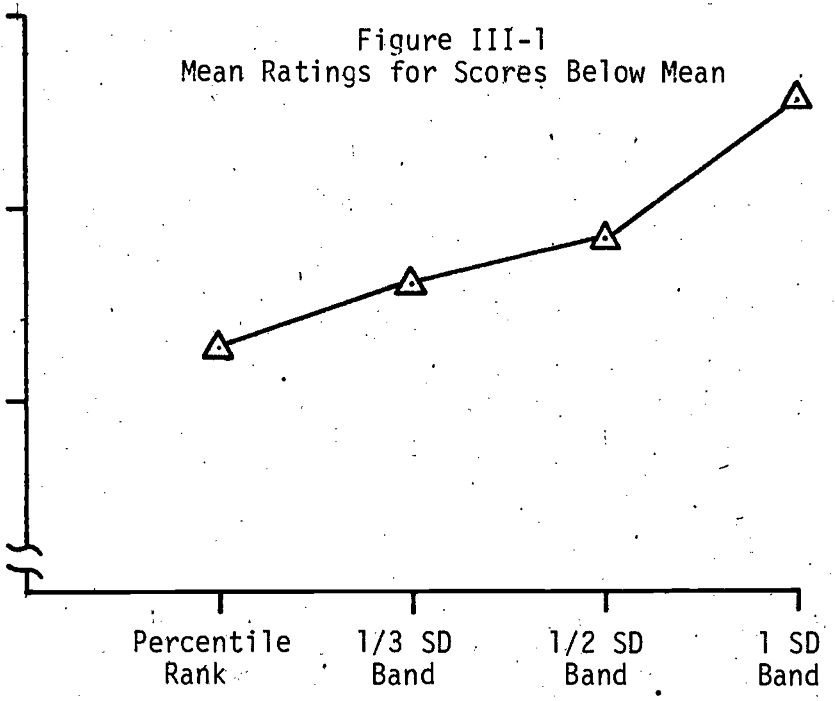

Figure III-2
Mean Ratings for Scores Above Mean

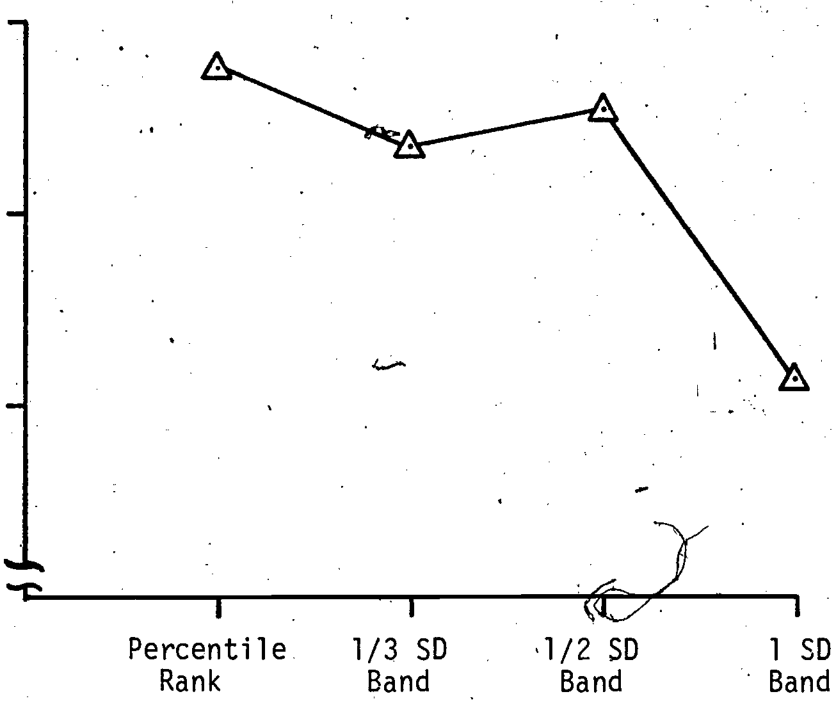

TABLE III-4
Summary of Tests of Trend for Scores Below Mean

Trend	df	MS	F	Probability
Linear	1	88.46	107.74	.000
Quadratic	1	4.40	5.36	.020
Cubic	1	2.14	2.60	.104
Residual	312	0.82		

TABLE III-5
Summary of Tests of Trend for Scores Above Mean

Trend	df	MS	F	Probability
Linear	1	109.71	110.49	.000
Quadratic	1	28.81	29.01	.000
Cubic	1	23.05	23.21	.000
Residual	312	0.99		

towards the mean. From these data, it is clear that trends in the interpretations of given scores can be identified, and the shape of the trend depends upon whether the scores are below or above the mean.

Summary

While the task in this study was contrived, the data suggest some very interesting conclusions may be drawn. First, educators -- if the sample used in this study is at all representative of teachers elsewhere -- do interpret point estimates and interval estimates differently. The general trend was to perceive a score as being closer to the mean when presented in increasingly wider interval estimates. In other words, these teachers tended to give systematic overestimates of scores below the mean and underestimates of scores above the mean when those scores were presented in interval band form. On the one hand, this is not unreasonable when the test reliability is zero, as the best point estimate for a randomly selected individual is the group mean. However, for the narrow and very narrow intervals, which represented reliabilities of .75 and .89, respectively, such an interpretation strategy is clearly inappropriate. This brings us to the second conclusion, that these teachers did not demonstrate an understanding of how a confidence band should be interpreted. Finally, since confidence bands better express the degree of accuracy with which human performance may be measured, reporting procedures may require a thorough examination if the producer wishes folks to draw appropriate interpretations from the data.

Substudy IV

Teachers' Estimates of Costs and Losses in Decisions

Moving from an interpretation to some definite action requires that a decision be made. The quality of a decision will depend upon the quality of information available for processing, as well as the capability of the individual to interpret and integrate the information in a sound manner. Thus, the perceived quality of information available to teachers will, apart from their skill at interpreting it, affect the kinds of decisions which teachers make. This conclusion is underscored by the work of Fleming and Antonen (1971), Goslin (1967); Rudman et al. (1980) and Kellaghan, Madaus and Airasian (1982). A second, personal factor which might affect behavior is the perception of how probable a correct decision might be. Finally, the perceived consequences or risks of an incorrect decision may well affect the choices which people make (Kahneman and Tversky, 1973).

The present ~~study was~~ initiated to answer three specific questions related to the quality of information and decision likelihoods: (a) How accurately do teachers believe standardized achievement test scores estimate pupil skill? (b) What outcomes or decisions are perceived of as having greater import? (c) What do teachers perceive to be the likelihood of making incorrect desisions? These questions serve to outline the focus of the study.

Methodology

Sample

Participants were 215 public school teachers from fourteen different school districts in Mississippi. These districts represented the western, southern, central and northeastern regions of the state. Approximately 80% of the sample were females and about 20% were males. These participants were in attendance at a week-long workshop on test development.

Instruments

Data for the first question come from a three-response task asking participants to judge the percent of students whose test scores on the California Achievement Test (CAT) represent an accurate reflection of their true skill; the percent who receive a too-low score; and the percent who receive a too-high score. The directions reminded the participants that these three values should sum to 100%. This measure is represented by items 1-3 of the booklet presented in Appendix F.

Data for question two came from one of two sets of six forced-choice stimuli for which participants were asked which of two outcomes they believed to be worse. These two sets differed in that one posed the question for atypical students (either very good or very poor), while the second posed the question for average students.

An example of the "loss ratio" items is:

Select the statement which you believe is the worse of the pair of statements.

Which is WORSE:

a. Accidentally placing a poor student in an advanced group or class.
b. Accidentally placing a good student in a remedial group or class.

The two sets of stimuli are contained in the first and second booklets in Appendix F. In each booklet, the forced-choice stimuli are items 4-9.

Data for the third question came from one of two sets of seven forced-choice stimuli for which participants were asked which of two outcomes they believed to be the more likely. These two sets differed in that one posed the question for atypical students (either very good or very poor), while the second posed the question for average students.

An example of the "likelihood" items is:

Select the statement which you believe is the MORE LIKELY of the pair of statement to occur. For each question, the student is of AVERAGE achievement level.

Which is MORE LIKELY:

a. A student performs very well on a classroom test.
b. A student performs very poorly on a classroom test.

The two sets of stimuli are combined in the first and second booklets in Appendix F. In each booklet, the forced-choice stimuli are items 10-16.

Participants completed the entire booklet in a single session. All participants were able to complete the three parts easily within twenty-five minutes.

Results

Question 1: How accurately do teachers believe standardized achievement tests estimate pupil skill?

Overall, the mean estimate for the percent of accurate scores was 62.6. Mean estimates of the frequency of too-low scores and too-high scores were 22.5 and 14.7, respectively. This suggests that these teachers believe that standardized achievement tests are on target about two-thirds of the time. Further, when the tests are believed inaccurate, the perceived tendency is to err towards an unrealistically low rather than unrealistically high score. A multivariate analysis of variance (MANOVA) was calculated to compare these estimated percentages among teachers of different gender, test course status, certification and teaching level.

The dependent variables chosen were the first two percentages (accurate and too-low). The reason for not including all three was the fact that forcing the values to sum to 100 introduces a dependency; that is, respondents only had two degrees of freedom in their selection. Independent variables included gender, test course status (whether or not participant had ever taken a course in tests and measurements); certification (A, representing a B.S. level; AA, representing an M.S. level; or AAA, representing an Ed.S. level of coursework); and teaching level (elementary, secondary or both). A summary of the main effects MANOVA contrasts is presented in Table IV-1.

In each case, there was no statistically significant difference among the contrasted groups. Interactions, not presented in the Table IV-1, were also not significant.

Thus, the perceived frequencies of right or wrong results coming from a specific achievement test were similar regardless of respondent gender, test course status, certification or teaching level.

Question 2: What outcome or decisions are perceived of as having greater import?

The results of the forced-choice instrument measuring perceived losses associated with incorrect decisions are summarized in Table IV-2. Each of the items forced a choice between a false positive (e.g., a student passing a test when he or she did not know the material) or a false negative outcome (e.g., a student failing a test when he or she did in fact know the material). The tabled percentages represent the frequency that a particular outcome was selected as worse.

In general, there was congruence between the observed percentages for the atypical and average student sets. The types of outcomes can be conveniently divided into two classes: Test results or decisions and instructional outcomes. The overall ratio of false positive (FP) to false negative (FN) selections (called a loss ratio) was markedly affected based on which class of outcomes was examined. For test results, the loss ratios were 0.46 for atypical students and 0.78 for average students. This suggests that the participants believed the worse outcomes for students to be test scores or decisions which underestimate rather than overestimate the true level of performance. From a study using seventh grade students. Morse (1977) found that students would tend to agree. Their perceived loss ratio was 0.47.

The resulting loss ratio for the instructional outcomes was quite different, though. For the atypical student item set, the resulting FP/FN value was 4.72, while for the average student set, the value was 10.40. The participants believed that false positive outcomes are considerably worse than false negative outcomes for students. That is, the teachers would apparently choose to err in the direction of holding the student back rather than pushing too quickly. The marked difference between the loss ratios for the atypical and average students suggests that the perceived disparity in FP and FN instructional outcomes is seen as more severe for average students. The loss ratio of the seventh grade students in Morse's study (1977) was not nearly as dramatic a departure from the test outcomes value, being 2.60.

Question 3: What do teachers perceive to be the likelihood of making incorrect decisions?

The results of the forced-choice instrument measuring judged likelihoods associated with incorrect outcomes or decisions are summarized in Table IV-3. For these items, the congruence between the judgments for the atypical and average student sets was much closer than for the loss ratio items. A similar pattern of different perceptions of test or performance versus instructional outcome likelihoods was noted, though.

The judged likelihoods of incorrect test or performance outcomes suggest that false negative outcomes are considered the more likely (FP/FN = 0.54 and 0.64 for atypical and average students, respectively). The picture reverses for instructional outcomes, in which false positive outcomes are judged to be for more common (FP/FN = 2.51 and 3.75 for atypical and average students, respectively).

The estimates of too-low and too-high test performance, discussed above in Question 1, give an independent check for test outcome likelihood. For those data, the likelihood ratio (FP/FN) was 0.65, which is congruent with the values obtained from the likelihood item sets.

The estimates from this instrument, as should be obvious, are of relative error likelihood, as opposed to absolute likelihood judgments. The task used in Question 1, though, was an absolute judgment task. That its results were congruent with the relative judgments from the forced-choice instrument suggests that switching to judgments of absolute likelihoods might not alter the relative error estimates.

Summary

Teachers apparently have at least a modicum of faith in standardized achievement tests, at least in the accuracy of the resulting scores. When incorrect results arise, they are perceived as being more often lower rather than higher than appropriate. Test theory suggests that, if necessary assumptions hold, errors of measurement are random rather than systematic (Lord and Novick, 1968). Perhaps this opinion reflects personal observations of some students being unable to perform well.

A more important conclusion is that false negative outcomes are perceived as less desirable than false positive outcomes for test decisions, yet for instructional outcomes, a false positive outcome is considered much worse than a false negative. These two observations suggest that there is a perception of test decisions somehow being independent of instructional decisions or outcomes. In other words, the link between tests as an example of controlled assessment and subsequent instructional decisions for pupils is either not perceived as important or is ignored. Either way, these data suggest an incoherent system: the preferred error for testing is to pass the student who doesn't have the skills but the error of choice for instruction is to hold back students who do have the requisite skills.

The tabulation of likelihood estimates again suggests that these teachers -- and teachers in general if this sample is at all representative of other teachers -- are operating in an incoherent system, as a Bayesian statistician would use the term (Novick and Jackson, 1974). In order to minimize overall "cost" or "loss" to a system, the appropriate strategy is to alter likelihoods of outcomes so that the products of loss ratios and likelihoods are at a minimum. Yet, these data suggest that the most costly, or the least desirable, decisions or outcomes are considered to be the most likely outcomes. (The reader should note that these are relative error rates being discussed and not absolute rates.)

One possible hypothesis is that the error which is observed most often is that which becomes judged as the more severe. If true, this hypothesis would serve to explain, in large part, the observed results. However, the patterns observed in the judgments suggest that an alternative hypothesis that generally incoherent decision-making schemes are in effect in education settings must also be considered as a possibility.

TABLE IV-1

Summary of MANOVA Contrasts on "Hit Rate" Estimates

Contrast	Wilks' Lambda	Approximate F	df	Probability
Gender	.959	2.84	2,132	.062
Test Course	.992	0.51	2,132	.601
Certificate	.979	0.69	4,264	.596
Level	.982	0.58	4,264	.674

TABLE IV-2
Teacher Loss Ratios
by Type of Student

A. For atypical students:	Worse Outcome	
Circumstance	False Positive	False Negative
1. Incorrect placement	41%	59%
2. Test performance	23%	77%
3. Moving on vs. remaining with material	85%	15%
4. Speed of presentation of material	76%	24%
5. Test performance; minimal P-F	31%	69%
6. Outcomes of incorrect instructional decisions	87%	13%

Summary: Test results or decision (1,2,5) : FP/FN = 0.46
Instructional outcomes (2,3,6) : FP/FN = 4.72

B. For average students:	Worse Outcome	
Circumstance	False Positive	False Negative
1. Incorrect placement	37%	63%
2. Test performance	47%	53%
3. Moving on vs. remaining with material	90%	10%
4. Speed of presentation of material	90%	10%
5. Test performance; minimal P-F	47%	53%
6. Outcomes of incorrect instructional decision	95%	5%

Summary: Test results or decision (1,2,5) : FP/FN = 0.78
Instructional outcomes (2,3,6) : FP/FN = 10.40

Note: Values for parts A and B are based on 143 and 72 respondents, respectively.

TABLE IV-3
Teacher Likelihood Estimates
by Type of Students

A. For atypical students:

Circumstance	Worse Outcome False Positive	 False Negative
1. Student work is atypical	.20	.80
2. CAT score is opposite expectation	.38	.62
3. CAT score is too far in same direction	.43	.57
4. Classroom test performance is opposite expectation	.30	.70
5. Semester grade is opposite expectation	.88	.12
6. Placement is incorrect	.55	.45
7. Classroom test P-F status is opposite expectation	.45	.55

Likelihood ratios: Test or performance (1-4,7) : FP/FN = 0.54
Instructional outcomes (5,6) : FP/FN = 2.51

B. For average students:

Circumstance	Worse Outcome False Positive	 False Negative
1. Student work is atypical	.53	.47
2. CAT score is opposite expectation	.10	.90
3. CAT score is too far in same direction	.16	.84
4. Classroom test performance is opposite expectation	.74	.26
5. Semester grade is opposite expectation	.90	.10
6. Placement is incorrect	.68	.32
7. Classroom test P-F status is opposite expectation	.42	.58

Likelihood ratios: Test or performance (1-4,7) : FP/FN = 0.64
Instructional outcomes (5,6) : FP/FN = 3.75

Note: Values for parts A, B based on 143 and 72 respondents, respectively.

Substudy V

Factors Influencing Teacher Estimation of Pupil Performance

Educational decision-making is governed in larae part by the dynamics of how teachers interpret the performance information available to them. Interpretations, once drawn, form the basis for aotion. The question of concern here is: What factors may be shown to affect interpretation and judgment of pupil performance?

Even though the now-infamous "Pygmalion" study (Rosenthal and Jacobson, 1968) was widely critiqued for its lack of experimental rigor, the question was raised as to whether teachers' judgments of pupil performance could be affected by inaccurate or unrelated information. Fleming and Antonen (1971), in an attempt to replicate the Pygmalion study, did observe that teachers did vary considerably in the degree of accuracy they were willing to ascribe to different types of test performance information. This perceived accuracy varied as a function of the degree of utility which was attributed to the particular type of performance information.

Teachers apparently temper their judgments of performance information accuracy on the basis of other information. Frederickson and Marchie (1966) noted that teachers asked to rate the validity of a given test score for one of their pupils, were much more likely to accept score information congruent with their prior beliefs than to accept incongruent score information. Leither (1976) found that teachers, even when asked to avoid all extraneous data, had a marked tendency to draw upon their knowledge of unrelated student background information in order to interpret test performance information.

Examples of the types of extraneous information which have been shown to affect teacher judgments are many. Perhaps one of the most widely-publicized is that of the pupil's name. Harari and McDavid (1973) found significant differences in teacher ratings of the same student work depending upon what name was attached to the work.

There is evidence to suggest that prior information does mediate decisions made on follow-up information. Shavelson, Caldwell and Izu (1977) noted that such decisions are determined in part by the congruence of the follow-up information with initial data, as well as the reliability of the information. Further, the Shavelson et al. study suggests that while perceptions of pupil capability or chances for success are more readily altered than pedagogical decisions, the types of pedagogical actions teachers report as best for a particular pupil do change as their perceptions of the pupil's capability changes.

Thus, if teachers' judgments do have an effect upon their behavior towards pupils, it is important to examine factors which may contribute to these judgments. The present study incorporated each of the factors suggested by prior research results. Extraneous information was represented by inclusion of the pupil's first name. Use of prior information was presented by requiring two separate decision points, the second coming after the presentation of follow-up information on the pupil. Congruence of information was incorporated by matching or mis-matching initial and follow-up information (each was either positive or negative). Reliability of information was represented by providing quite different sources of information. In this way, the study allowed the examination of the interactive effects of these factors as they affected teacher decisions.

The specific questions which define the scope of the study were: (a) Does pupil gender, initial appraisal, follow-up information or reliability of information affect teacher decisions? (b) Are these differences in teacher decisions attributable to gender, training, certification or teaching level?

Methodology

Sample

Participants were 163 teachers, with varying levels of experience. Education levels were approximately evenly divided between undergraduate training only (46%) and graduate degrees (M.S. 43%, Specialist degree, 11%). All respondents were participants in a training workshop and voluntarily completed the instrument. Complete data were obtained from 157 of the 163 teachers (96%).

Instruments

The instrument used for this study was a slightly altered version of that used in the Shavelson et al. study. Respondents were presented with initial information for a "student" and were then asked to judge the chances (between 0 and 100%) of the student obtaining all A's ~~and~~ B's on the report card. The initial information varied by valence (either positive or negative in the description of pupil's ability, study habits and family background) and by gender (the student was given either a male or female name, no surname supplied).

After judging the child's chances for success, the follow-up information was presented. This information varied by valence (either positive or negative in the descriptions of the child's achievement and "attitude" towards school), and reliability (the information coming from reliable and authoritative sources or from unreliable and unauthoritative sources). Respondents were then asked to judge again, in light of the follow-up information, the child's chances for success in school.

The possible conditions thus formed a 2x2x2x2 factorial design, with the additional repeated measures dimension of teacher judgment. The factors were: initial information valence; pupil gender; follow-up information valence; and follow-up information reliability.

Other questions posed at the time of each judgment included: (a) whether the textbooks to be chosen for the student should be at, at or above, or below the student's grade level; (b) how the teacher would react if the child hesitated in answering a question in class; and (c) how important it was to praise the child every time he or she did good work. These questions are referred to as the textbook, questioning and reinforcement decisions, respectively. A copy of the information paragraph types and a sample booklet are contained in Appendix G.

The ordering of factor conditions was randomized prior to distribution of the booklets. Each participant was able to complete the task easily within twenty-five minutes.

Results

Question 1: Does pupil gender, initial appraisal, follow-up information or reliability of information affect teacher decisions?

Contrasts of initial judgments by valence and gender indicated a significant information valence effect, but only trivial differences due to pupil gender (F for valence = 268.91; F for gender = 0.86; df = 1/153). These results are displayed in Table V-1. Because of this, the initial judgments were used as a covariate for the contrasts of follow-up judgment by all four factors. The summary information for the ANCOVA contrasts of final probability estimates is contained in Table V-2.

From the data in Table V-2, it is apparent that when follow-up judgments are adjusted for initial judgments, pupil gender and follow-up information valence were significant main effects ($p = .011$ and $p < .001$, respectively). The reliability of the follow-up information, while not significant as a main effect, was part of significant two-way interactions with both initial and follow-up information valence ($p = .002$ and $p < .001$, respectively). No other interaction was statistically significant.

Means for the differences in judgment (follow-up estimate - initial estimate) by valence condition for male and female names are presented in Table V-3. These means suggest several important results: (a) When the two information sets were congruent in valence, the differences were considerably smaller than when they were incongruent. (b) Respondents were systematically favoring the male student over the female in their judgment revisions. Positive mean changes in judgment

TABLE V-1

ANOVA Summary of Initial Probability Estimates

Source	df	MS	F	Probability
Gender (G)	1	277.49	0.86	.355
Initial information valence (I)	1	86210.68	268.91	.000
G x I	1	660.40	2.06	.149
Residual	153	320.58		
Total	156	873.07		

TABLE V-2

ANCOVA Summary of Final Probability Estimates

Source	df	MS	F	Probability
Covariate	1	15799.35	48.60	.000
Gender (G)	1	2171.70	6.68	.011
Initial information valence (I)	1	671.70	2.07	.153
Follow-up information valence (F)	1	38294.21	117.81	.000
Follow-up information reliability (R)	1	89.15	0.27	.601
G x I	1	789.28	2.43	.121
G x F	1	84.31	0.26	.611
G x R	1	187.81	0.58	.488
I x F	1	166.74	0.51	.475
I x R	1	3277.68	10.08	.002
F x R	1	18187.49	55.95	.000
G x I x F	1	19.34	0.06	.808
G x I x R	1	82.28	0.25	.616
G x F x R	1	94.56	0.29	.591
I x F x R	1	363.28	1.12	.292
G x I x F x R	1	14.663	0.45	.832
Residual	140	325.06		
Total	156	888.10		

TABLE V-3

Mean Difference in Judgment by Information Valence and Gender

		Female Student				Male Student	
		Follow-up information				Follow-up information	
		-	+			-	+
Initial information	-	-12.10 (13.31)	28.57 (32.41)	Initial information	-	0.65 (19.63)	35.29 (28.96)
	+	-30.46 (28.16)	2.24 (5.95)		+	-24.10 (26.25)	3.10 (10.54)

Note: Values in parentheses are standard deviations.

were larger for the male student, while negative mean changes were larger for the female student. The size of this difference by gender ranged from 0.86 percentage points for the dual positive valence to 12.75 for the dual negative valence. In fact, for the dual negative valence, judgments for males were revised up about two-thirds of a point while the mean judgment for females declined by over twelve points.

Summary statistics for the significant two-way interaction variables are contained in Table V-4. The patterns for initial and follow-up data valences were congruent. The reliable follow-up information resulted in differences about six or more times as large as those for the unreliable follow-up information. For the incongruent data valences, though, the reliable follow-up information resulted in differences slightly over three times as large as those for unreliable information.

Because of the unexpected impact which pupil gender had on the outcomes, a follow-up study was planned. In this study, a total of fifty-six public school teachers from two school districts in southwestern Mississippi was asked to participate. The design of the follow-up study was essentially the same, with two differences. First, two female names, Carol and Susan were used instead of a male and female name. Second, only reliable follow-up information was presented. Thus, the study represented a 2x2x2 design of name by initial information valence by follow-up information valence. Fifty-four usable booklets were turned in.

The results, presented as an ANCOVA contrast of final probability estimates using initial estimates as the covariate, are summarized in Table V-5. No significant main effect other than follow-up valence was observed ($F = 94.15$; $p < .001$). None of the interactions was statistically significant. Thus, the observed differences due to gender in the main study were apparently not due to selection of a disagreeable female name. Whether it was caused, in part, by an especially fortuitous choice of male name is still open to question.

Path analysis models were generated and tested for each of the three decisions called for: textbook, questioning and reinforcement. Following the Shavelson et al. approach, two interaction variables, SV_1 and RV_2 were created. SV_1 represents interaction of gender and initial valence, while RV_2 represents the interaction of reliability and follow-up valence. However, the SV_1 variable was not a significant contributor to either initial prediction (PE_1) or resulting decision (TD_1, QD_1 or RD_1), as suggested by the results in TABLE V-2. The valence of initial information (V_1) was used as the sole exogenous variable for initial prediction, while gender (G) was used as one of the two purely exogenous variables for the follow-up prediction (PE_2). Kenny (1979) outlines the mechanics of generating and testing path models.

TABLE V-4

Mean Differences in Judgment by Information Valence and Reliability

		Reliable Information:				Unreliable Information:	
		Follow-up information				Follow-up information	
		-	+			-	+
Initial information	-	-12.95 (14.35)	50.83 (33.00)	Initial information	-	0.83 (18.09)	14.25 (13.89)
	+	-42.10 (28.81)	6.70 (8.09)		+	-13.62 (16.41)	-1.19 (6.88)

Note: Values in parentheses are standard deviations.

TABLE V-5

ANCOVA Summary of Final Probability Estimates from Follow-up Study

Source	df	MS	F	Probability
Covariate	1	2530.07	6.31	.016
Name (N)	1	6.53	0.02	.899
Initial information valence (I)	1	444.97	1.11	.298
Follow-up information valence (F)	1	37767.23	94.15	.000
N x I	1	649.31	1.62	.210
N x F	1	601.12	1.50	.227
I x F	1	703.586	1.75	.192
N x I x F	1	445.97	1.11	.297
Residual	45	401.13		
Total	53	1144.30		

Figures V-1, V-2 and V-3 illustrate the path models which represent the best fit to the sample data for the textbook decision, the questioning decision and the reinforcement decision, respectively. For the initial textbook decision, it is apparent that the initial information valence effect overshadows that of the initial prediction by a ratio of about two to one. The subsequent information, contained in the RV_2 variable, the gender (G) and the initial prediction all combine to affect the follow-up prediction of success (PE_2). As a result, the follow-up decision is affected most strongly by the follow-up prediction, followed by the RV_2 variable and the initial text decision (TD_1). In this instance, both prior and collateral information are being combined in the probability estimates and subsequent decision.

The same cannot be said to hold for questioning strategy decisions. As illustrated in Figure V-2, the choice of questioning strategy is apparently unaffected by any variable other than initial questioning decision (QD_1). In other words, questioning strategy was essentially invariant across the observed factors; teachers seem to have a preferred style or strategy and choose not to alter it.

Reinforcement strategy decisions, though, were affected to a degree by follow-up information. Figure V-3 illustrates that for both initial and follow-up decisions (RD_1, RD_2), the prediction estimates and purely exogenous variables all combined to affect the decision.

Question 2: Are there differences in teacher decisions attributable to gender, training, certification or teaching level?

An analysis of covariance, using initial probability estimates of success as the covariate, was calculated in order to contrast the various levels of the personal variables considered. The ANCOVA results are presented in Table V-6. As is suggested by the figures in Table V-6, none of the main effects examined -- gender, whether or not coursework in tests and measurements had been taken, level of certification (A, AA or AAA) or teaching level (elementary, secondary or both) -- made a difference in the adjusted final probability estimates. Because some of the two- and three-way interaction cells were empty for this sample, only main effects were examined, and a pooled within-cell variance estimate was used.

Summary

That teachers' judgments depend upon certain factors is apparently a reasonable proposition. Teachers' responses in this study suggest that they are sensitive to the congruence of new information with prior information, the reliability of information, the gender of the student and the valence of performance information. Why male names should be

Figure V-1
Path Model and Coefficients for Text Decision

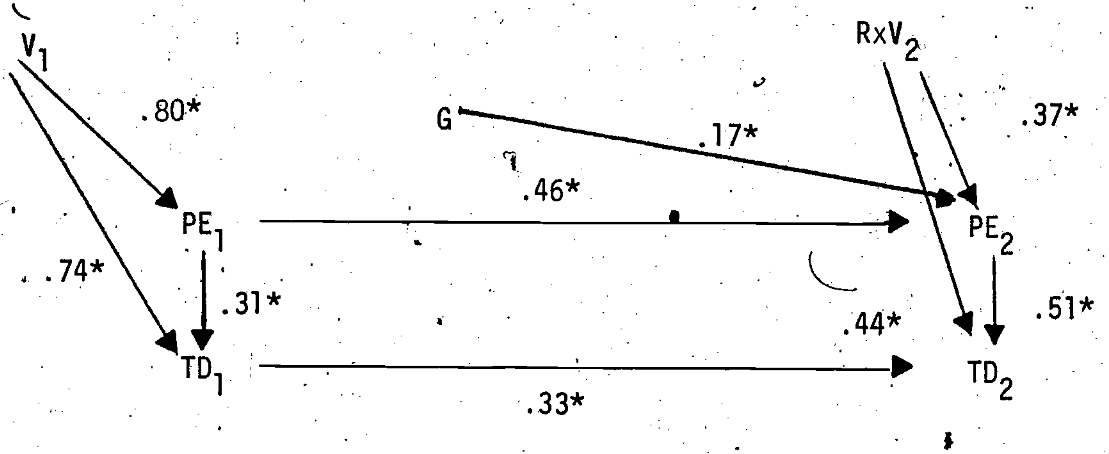

Figure V-2
Path Model and Coefficients for Questioning Decision

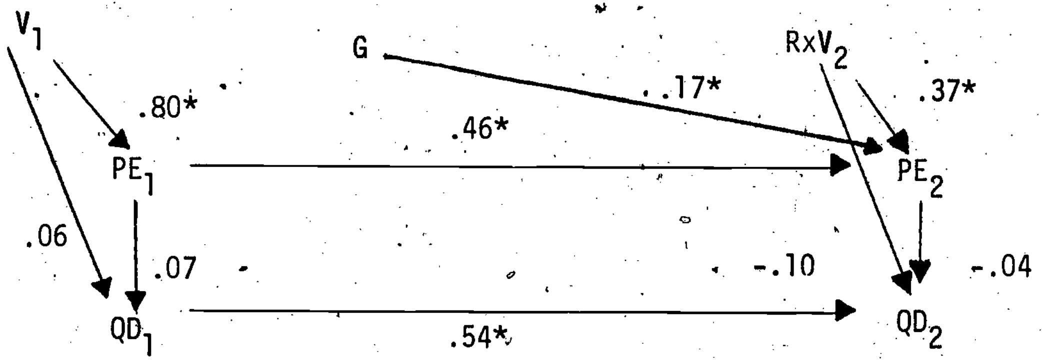

Figure V-3
Path Model and Coefficients for Reinforcement Decision

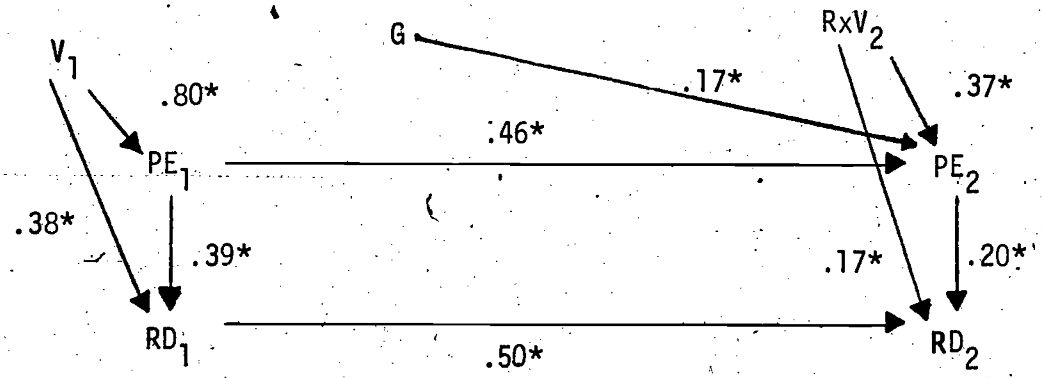

TABLE V-6

ANCOVA Summary of Final Probability Estimates
For Teacher Groups

Source	df	MS	F	Probability
Covariate	4	15016.10	33.02	.000
Gender	1	369.14	0.81	.369
Test Course	1	201.24	0.44	.507
Certification	2	454.47	1.00	.371
Level	2	1087.64	2.39	.095
Residual	139	454.69		
Total	149	877.96		

systematically favored over female names is not clear. However, should there be even the slightest link between this difference observed on an artificial task and behavior in the classroom, then serious consideration should be given to approaches by which such inequities may be reduced. Second, as Shavelson et al. suggest, it appears that teachers' decision-making -- with the exception of questioning strategy -- is somewhat Bayesian in nature. Unfortunately, the participants in this sample were not equitably Bayesian. Third, the link between teacher perception of pupil capability and subsequent behavior towards the students deserves further investigation. Fourth, teacher characteristics do not appear to have any systematic effect on judgments of a student's chances for "success" in school. These results indicate that both relevant and irrelevant information are incorporated in decision-making. One possible implication is that teacher preparation should include training on sound decision-making.

Substudy VI

Changes in Teachers' Knowledge and Perceptions of Test Score Data

That teachers' knowledge and perceptions of test score data are perhaps not as sound as desirable is a fairly common conclusion (Hastings et al., 1960; Goslin, 1978; Rudman et al., 1980). However, as Rudman and colleagues pointed out in their 1980 review of assessment practice, not much has been done to investigate whether teachers' knowledge can be changed by direct intervention, such as staff development training sessions (e.g., in-service education). The results of the previous substudies contained in this project report suggest that no systematic differences in perceptions, use or interpretation of performance data could be attributed to whether or not the participant had taken one or more courses in tests and measurements. It may well be the case that the topics traditionally covered in such courses emphasize statistical concepts and treat the topics of interpretation and use only lightly, if at all. A second possibility is that the time separating the course work from the present is simply too great to allow the retention of measurement concepts.

The present study was initiated to provide insight on two specific questions: (a) Can teachers' knowledge or perceptions of test score data be changed as a result of short-term, directed training? (b) Are there differences in the degree of this change attributable to teachers' measurement background, certification or teaching level?

Methodology

Sample

Participants in the study were 245 public school teachers from a Mississippi school district. The school district offers instruction from grades one to twelve. The racial mix of the teachers was about 75% white and 25% black. By gender, the percentage of females was about 75%, that of males about 25%. The teachers within the system represented a variety of teacher training institutions attended. The town in which the school system is based has a population of slightly over 15,000, making it a medium-size city for Mississippi.

Instruments

Four separate subtests were used in this study, two of which related to perceptions of test score data while the others measured knowledge of tests and test scores. These subtests are discussed individually in detail below. A copy of all instruments is contained in Appendix H.

Validity subtest. This subtest was the same as that used for the first research question in Substudy I. It was comprised of one of two sets of eight items, each of which gave test and nontest information for a hypothetical student. The participant was then asked to judge the given test score as being valid, questionable or invalid. An example from the validity judgment subtest is:

> A female student, sixth grade, average grade of D in reading and social studies. New CAT-77 reading comprehension percentile is 92. This score is:
> a. Valid
> b. Questionable
> c. Invalid

A response of "Valid" was coded as three points, a rating of "Questionable" as two and "Invalid" as one point. The ratings were then summed across items. Thus, high scores represent a greater perceived validity of test scores -- in both congruent and incongruent settings -- while low scores represent a lower degree of perceived validity. Overall internal consistency reliability of this measure was .80.

Knowledge subtest. This subtest was designed to assess how well participants could interpret both classroom and standardized achievement test score data. Several of the ten items came from those used in surveys of test coordinators' (Morse, 1978) and accountability coordinators' (Hills, 1977) perceptions of teacher competence in measurement. Two of the items came from the Newman and Stallings (1980) study. Several others came from a course in measurement taught by the author. All items had been thoroughly pretested. As used in the present study, the raw score (number right) was the criterion variable. Overal internal consistency reliability was .65, which compares favorably with values reported by Hastings et al. (1960) and Newman and Stallings for much longer tests.

Test-wiseness subtest. Understanding of sound item and test construction practice should permit an examinee to detect and take advantage of poorly constructed tests. In addition, test-wiseness is a trait which has been shown to be trainable (Morse and Morse, 1980) for both those skills from the Millman, Bishop and Ebel (1965) hierarchy which are independent of the test constructor and those dependent upon the test constructor. The set of fourteen items was drawn from a study in which Morse (1980) found that the test-wiseness skills dependent upon test and test constructor were significantly more difficult to apply successfully than were the skills independent of test and test constructor (the population used in that study was fifth and sixth grade students). The selected items were therefore evenly balanced for test-dependent and test-independent skills. An example of a test-wiseness item is:

If something is inflammable, it will
a. resist burning
b. not catch on fire
c. not be consumed by flames
*d. easily ignite

The skill required here is to avoid selection of responses a, b and c since they each imply a similar result. Choice d, being unique, is the preferred selection. Simple raw score was used. Overall internal consistency reliability of this measure was .77.

Preference subtest. This subtest was taken from the card sort task used by Goslin (1967) and originally used by Hastings et al. (1960). It is comprised of twenty-eight "records" each containing some combination of test and nontest information for a hypothetical student. For each case, the participant was asked to judge whether the high-school student should be placed in a regular or advanced science class. On fourteen of the cards, the data were uniformly positive or negative, which should lead to little variation in assignment. On the remaining fourteen, however, the information was incongruent (often the record was incomplete, also). Thus, the chosen assignment could be an indicator of the degree to which the participant attended to the test information as opposed to the nontest information.

Two modifications were made for this study concerning the subtest. First, participants only were given fifteen of the cases to assign. Two different forms were prepared, each having two common and thirteen unique cases. Forms were then randomly assigned to the participants. Subsequent examination of the two "anchor" items indicates no systematic differences in responses could be ascribed to the form received. The second difference is that the scoring procedure used in the Goslin study was altered slightly. The final score, though, still represented a relative percentage of preference of test versus nontest information. Hence, scores over 50 indicate a more frequent dependence upon the test data, while scores below 50 represent a more frequent use of the nontest data in making the assignment. Internal consistency reliability for this scale was .85. The items and scoring procedure are contained in Appendix H.

Design

The research design was a hybrid quasi-experimental approach. Using a modified Campbell and Stanley (1966) notation, the design was:

				Fifteen-day span
A	O_1	X	O_2	
A	O_1	X		O_2

where: A indicates random assignment by school (9 in all) to group
X denotes training in test interpretation
O_1 denotes pretest
O_2 denotes immediate or delayed posttest.

Logistic considerations prevented true random assignment, which would have been preferable. Also, the retention test could not be spaced as far behind the treatment as proposed due to school calendar limitations. The design did allow for inferences as to pre-instructional level, pre- to post-instructional changes and short-term retention.

Treatment

The treatment, part of an on-going program in test development, consisted of a single three-hour afternoon session. The general topics which were covered emphasized interpretation of test score data rather than attitudes toward tests and test data. An outline of the session follows.

Training Session Outline

"Using Tests and Test Scores Wisely"

I. Introduction (20 minutes)
   A. The many uses of test results
   B. The many types of tests
   C. The many types of test scores
II. Comparing different test scores (40 minutes)
   A. Raw scores vs. derived scores
   B. Common derived scores and their interpretation
   C. Appropriate scores for norm-referenced, criterion-referenced tests
III. Fallibility of test scores (30 minutes)
   A. Errors of measurement
   B. Confidence bands
   C. Factors which affect accuracy of test scores
BREAK (20 minutes)

IV. Small group sessions (by grade level) (45 minutes)
   A. Interpreting standardized test results
   B. Interpreting criterion-referenced and classroom test results
   C. Sample student records--interpretation of test scores and contrast with other achievement data
V. Summary and discussion (25 minutes)

Results

Question 1. Can teachers' knowledge or perceptions of test score data be changed as a result of short-term, directed training?

Summary statistics for all subtests on each occasion are presented in Table VI-1. An increase in the overall mean scores was noted on each of the subtests. Because no differences were observed between the immediate and short-term retention outcomes, those results were pooled. It may well be the case that there was considerable mental note-exchanging taking place between the two groups, contaminating the results to an indeterminate extent.

Simple pretest-posttest contrasts indicated statistically significant gains for the knowledge subtest ($F = 264.95$; $df = 1,244$; $p < .001$) and for the test-wiseness subtest ($F = 64.92$; $df = 1,244$; $p < .001$). However, there were no systematic differences for either the validity subtest ($F = 0.52$; $df = 1,244$) or the preference subtest ($F = 0.34$; $df = 1,244$). These differences, expressed as effect sizes, are summarized in Table VI-2. The net change for the knowledge subtest was about a full standard deviation, while that for the test-wiseness subtest was about one-half a standard deviation.

Intercorrelations among the subtests at each occasion are presented in Table VI-3. It is interesting to note that the values changed only modestly from the first test to the follow-up test.

Question 2: Are there differences in the degree of this change attributable to teachers' measurement background, certification or teaching level?

A multivariate analysis of variance (MANOVA) was calculated for the subtest vector comparing the various teacher characteristics on the pre-instructional test. These results are presented in Table VI-4. The main effect of teaching level (elementary, secondary or both) was significant at the .05 level, as was the certificate by level interaction and the course by certificate by level interaction. Univariate ANOVA contrasts were then calculated for each significant effect. These results are contained in Table VI-5. Only one contrast for each interaction was statistically significant at the .05 level. For the certificate by teaching level interaction, the difference was observed on the knowledge subtest. Cell means and sizes are

TABLE VI-1

Summary Statistics for
Knowledge and Perception Measures

	Initial Test				Follow-up Test	
Measure	Mean	S.D.	Alpha	K	Mean	S.D.
Validity	17.44	8.27	.80	8/8	17.72	8.52
Knowledge	3.96	1.69	.65	10	6.19	2.53
Test-wiseness	9.91	2.78	.77	14	11.37	2.88
Preference	64.63	21.35	.85	5/9	65.10	21.48

Note: Values based on 245 respondents.

K represents number of scored items; dual values indicate alternate forms.

TABLE VI-2

Overall Effect Sizes for Change in Knowledge and Perception Scores

Validity	Knowledge	Test-wiseness	Preference
.03	1.04	0.51	.02

Note: Effect size is $(\bar{X}_2 - \bar{X}_1)/S$ pooled; values based on 245 respondents.

TABLE VI-3

Product-moment Correlations Among
Knowledge and Perception Scores

	Validity	Knowledge	Test-wiseness	Preference
Validity	-	-.33	-.27	-.37
Knowledge	-.30	-	.21	.20
Test-wiseness	-.29	.21	-	.57
Preference	-.36	.22	.53	-

Note: All values based on 245 respondents.

Upper diagonal values are initial test results; lower diagonal values are second test results.

TABLE VI-4

Summary of MANOVA Contrasts of Teacher Groups on Initial Knowledge and Perception Scores

Contrast	Wilks' Lambda	Approximate F	df	Probability
Test Course (T)	.997	0.13	4,205	.971
Certificate (C)	.967	0.86	8,410	.550
Level (L)	.925	2.05	8,410	.040
T x C	.978	0.57	8,410	.801
T x L	.976	0.64	8,410	.748
C x L	.875	1.75	16,627	.034
T x C x L	.920	2.19	8,410	.027

TABLE VI-5

Selected Univariate ANOVA Contrasts of Teacher Groups on Initial Knowledge and Perception Scores

Contrast = Level (df = 2,208):

Variable	Hypothesis MS	Error MS	F	Probability
Validity	15.39	5.59	2.75	.066
Knowledge	7.36	2.48	2.96	.054
Test-wiseness	0.04	7.21	0.01	.995
Preference	460.53	346.92	1.33	.267

Contrast = Certificate by Level (df = 4,208):

Variable	Hypothesis MS	Error MS	F	Probability
Validity	9.49	5.59	1.70	.152
Knowledge	6.90	2.48	2.78	.028
Test-wiseness	10.92	7.21	1.51	.200
Preference	630.82	346.92	1.82	.127

Contrast = Test Course by Certificate by Level (df = 2,208):

Variable	Hypothesis MS	Error MS	F	Probability
Validity	42.28	5.59	7.56	.001
Knowledge	1.39	2.48	0.56	.573
Test-wiseness	4.87	7.21	0.68	.510
Preference	41.81	346.92	0.12	.887

presented in Table VI-6. The primary "cause" of the interaction appears to be the difference among the AA (M.S. or M.Ed.) level certificate holders' pattern of means. Specifically, for elementary teachers, the average is lower than that of the A certificate holders, but is higher for secondary teachers.

Table VI-7 presents cell means and sample sizes for the three-way interaction as the validity subscale. The relatively small numbers of teachers in the no test course group raises a question as to how stable this interaction might be. The flip-flop between higher and lower means across the certificate and teaching level combinations explains why the interaction was significant; there does not appear to be any clear-cut pattern, though, to this interaction.

Results on the follow-up test scores adjusted for the initial differences, again comparing the teacher characteristic groups, are presented in Table VI-8. The MANCOVA results indicate that none of the main effects or interactions was statistically significant. Hence, follow-up univariate tests were not calculated.

Certain combinations of teacher characteristics did serve to explain part of the initial differences observed on the validity and knowledge subtests, but were not systematically related to the degree of change on the set of subtests.

Summary

A short-term training session can effect significant gains in teacher knowledge of interpretations of test scores as well as in measured test-wiseness. No decrement in performance was observed when teachers tested immediately after the instruction were compared with teachers tested fifteen days later. No changes were observed on the two perception subscales, which is not surprising since the focus of the training was on cognitive rather than affective outcomes.

There were initial differences in knowledge of score interpretation and perceived validity of test score data due to combinations of certificate level, teaching level and measurement course work status. When follow-up scores were adjusted for initial scores, no systematic differences among the various combinations of teacher characteristics were observed.

The implications of this study are important for future research endeavors. First, presence or absence of measurement course work does not appear to make much difference in knowledge or perceptions of test score data. Perhaps elapsed time, unrelated content or a combination of the two could explain why those teachers having had

TABLE VI-6

Certificate by Level Means for Initial Knowledge Scores

	Level		
Certificate	Elementary	Secondary	Both
A	3.98 (43)	3.99 (70)	3.50 (6)
AA	3.54 (39)	4.71 (35)	7.00 (1)
AAA	4.06 (17)	4.73 (11)	2.50 (2)

Note: Numbers in parentheses are cell sizes.

Table VI-7

Test Course by Certificate by Level Means for Validity Subtest

	Teaching Level					
Certificate	Elementary		Secondary		Both	
	Course	No Course	Course	No Course	Course	No Course
A	17.35 (34)	15.78 (9)	15.85 (46)	15.88 (24)	15.67 (3)	13.00 (3)
AA	15.59 (32)	19.71 (7)	16.21 (28)	16.00 (7)	14.00 (1)	--
AAA	16.86 (14)	15.00 (3)	16.00 (9)	18.50 (2)	14.00 (2)	--

Note: Numbers in parentheses are cell sizes.

TABLE VI-8

Summary of MANCOVA Contrasts of Teacher Groups on Follow-up Knowledge and Perception Scores

Contrast	Wilks' Lambda	Approximate F	df	Probability
Test Course (T)	.968	1.69	4,201	.155
Certificate (C)	.964	0.92	8,402	.501
Level (L)	.973	0.70	8,402	.689
T x C	.965	0.90	8,402	.514
T x L	.963	0.97	8,402	.461
C x L	.923	1.02	16,615	.432
T x C x L	.973	0.69	8,402	.701

measurement training performed no differently than those without. Second, the cognitive skills related to knowledge of test data interpretation and test-wiseness can be improved as a result of a modest intervention. For the authors of studies which suggest that sound understanding of the relevant principles is a necessary precondition to sound decision-making, these results should be encouraging. Finally, school systems should consider the possibility of devoting at least some in-service time to enhancement of teachers' skills in the interpretation of test score data. This is one of the few examples of a policy from which virtually everyone -- not the least important being the child about whom the decisions are being made -- stands to benefit.

References

Anastasi, A. Psychological Testing (4th ed.). New York: MacMillan, 1976.

Beggs, D. L., Mayer, G. R., & Lewis, E. L. The effects of various techniques of interpreting test results on teacher perceptions and pupil achievement. Measurement and Evaluation in Guidance, 1972, 5, 290-297.

Buros, O. K. (Ed.). The eighth mental measurement yearbook. Highland Park, NJ: The Gryphon Press, 1978.

Burry, J., Catterall, J., Choppin, B., and Dorr-Bremme, D. Testing in the nation's schools and districts: How much? What kinds? To what ends? At what costs? CSE Report No. 194. Los Angeles: Center for the Study of Evaluation, University of California, 1982.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Farr, R., & Griffin, M. Measurement gaps in teacher education. Journal of Research and Development in Education, 1973, 7, 19-28.

Fleming, E. D. & Antonen, B. Teacher expectancy or my fair lady. American Educational Research Journal, 1971, 8, 241-252.

Frederickson, R. H., & Marchie, H. E. Teachers view test results. The Clearing House, 1966, 40(6), 357-358.

Goslin, D. A. Teachers and testing. New York: Russell Sage Foundation, 1967.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hambleton, R. K. & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Harari, H., & McDavid, J. W. Name stereotype and teachers' expectations. Journal of Educational Psychology, 1973, 65, 222-225.

Hastings, J. T., Runkel, P. J., Damrin, D. E., Kane, R. B., & Larson, G. L. The use of test results. Cooperative Research Project No. 509. Urbana, IL: Bureau of Educational Research, University of Illinois, 1960.

Hills, J. R. Coordinators of accountability view teachers' measurement competence. Florida Journal of Educational Research, 1977, 19, 39-44.

Hills, J. R. Measurement and evaluation in the classroom (2nd ed.). Columbus, Ohio: Merrill, 1981.

Hoepfner et al. (Eds.). CSE-RBS test evaluations: Tests of higher-order cognitive, affective, and interpersonal skills. Los Angeles, CA: University of California at Los Angeles, Center for the Study of Evaluation, 1972.

Kahneman, D., & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. The effects of standardized testing. Boston, MA: Kluwer-Nijhoff Publishing, 1982.

Kenny, D. A. Correlation and causality. New York: Wiley, 1979.

Leiter, C. K. W. Teachers' use of background knowledge to interpret test scores. Sociology of Education, 1976, 49, 59-65.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Millman, J., Bishop, C. H., & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement. 1965, 25, 707-726.

Morse, D. T. Assessment of selected efficient measurement strategies in an individualized curriculum. Florida Journal of Educational Research, 1977, 19, 18-24.

Morse, D. T. How Mississippi district test coordinators perceive classroom teachers' measurement competence. A paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, 1978.

Morse, D. T. The relative difficulty of selected test-wiseness skills. A paper presented at the annual meeting of the National Council on Measurement in Education, Boston, 1980.

Morse, L. W., & Morse, D. T. The effects of teacher style and specificity of instructional materials on learning and retaining different tasks. A paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Morse, P, K. Reporting test results: Percentile bands vs. percentile ranks. *Journal of Educational Measurement*, 1964, *1*(2), 139-142.

Newman, O. C., & Stallings, W. M. Teacher competency in classroom testing, measurement preparation, and classroom testing practices. A paper presented at the annual meeting of the National Council on Measurement in Educat.on, New York, 1982.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

Pipho, C. (Ed.). *Phi Delta Kappan*, 1978, *59*(9).

Rosenthal, R., & Jacobson, L. *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart and Winston, 1968.

Rudman, H. C., Kelly, J. L., Wanous, D. S., Mehrens, W. A., Clark, C. M., & Porter, A. C. *Integrating assessment with instruction: A review (1922-1980)*. Research Series No. 75. East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1980.

Shavelson, R. J., Caldwell, J., & Izu, T. Teachers' sensitivity to the reliability of information in making pedagogical decisions. *American Educational Research Journal*, 1977, *14*, 83-97.

Shoemaker, D. M. *Use of sampling procedures with the USOE Title I evaluation models*. (Monograph No. 4 in USOE Office of Evaluation and Dissemination series) Washington, D. C.: U. S. Government Printing Office, 1979.

Tallmadge, G. K. *The joint dissemination review panel ideabook*. Mountain View, CA: RMC Research Corporation, October 1977. (Report funded by the National Institute of Education under Contract #NIE-IA-7706).

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects (Revised). Los Altos, CA: RMC Research Corportion, December, 1974 (Technical Report No. UR-240).

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Appendix A

Validity Judgment Items

Key to Validity Judgment Items

Set	Item	Child's Gender	Child's Grade	Nontest Performance	Test Score
1	1	M	8	Low	High
	2	F	10	High	Low
	3	F	3	High	Low
	4	M	4	High	High
	5	M	11	Low	Low
	6	F	6	Low	High
	7	M	7	High	Low
	8	F	2	High	High
2	1	M	5	Low	High
	2	F	2	High	Low
	3	F	12	High	Low
	4	M	9	High	High
	5	M	6	Low	Low
	6	F	9	Low	High
	7	M	1	Low	Low
	8	F	8	High	High

Set 1

For items 1-8, decide whether the newly received test score is valid, questionable, or clearly invalid.

1. A male student, eighth grade, average grade of C-. New IQ score is 130. This score is:

   a. Valid
   b. Questionable
   c. Invalid

2. A female student, tenth grade, average grade of B+. New IQ score is 82. This score is:

   a. Valid
   b. Questionable
   c. Invalid

3. A female student, third grade, average grade of A-. New CAT-77 reading NCE is 36. This score is:

   a. Valid
   b. Questionable
   c. Invalid

4. A male student, fourth grade, average grade of 86 in mathematics. New CAT-77 math percentile is 90. This score is:

   a. Valid
   b. Questionable
   c. Invalid

5. A male student, eleventh grade, average grade of 74 in English. New CAT-77 language arts percentile is 38. This score is:

   a. Valid
   b. Questionable
   c. Invalid

6. A female student, sixth grade, average grade of D in reading and social studies. New CAT-77 reading comprehension percentile is 92. This score is:

a. Valid
b. Questionable
c. Invalid

7. A male student, seventh grade, average grade of A in mathematics. New CAT-77 mathematics concepts and problem solving percentile is 28. This score is:

a. Valid
b. Questionable
c. Invalid

8. A female student, second grade, average grade of "excellent" in reading. New CAT-77 reading vocabulary percentile is 78. This score is:

a. Valid
b. Questionable
c. Invalid

Set 2

For items 1-8, decide whether the newly received test score is valid, questionable, or clearly invalid.

1. A male student, fifth grade, average grade of 78 in English. New CAT-77 language arts percentile is 93. This score is:

a. Valid
b. Questionable
c. Invalid

2. A female student, second grade, average mark on report card is 90. New IQ score is 82. This score is:

a. Valid
b. Questionable
c. Invalid

3. A female student, twelfth grade, with a semester average of 93 in Senior English. New CAT-77 language arts percentile is 30. This score is:

   a. Valid
   b. Questionable
   c. Invalid

4. A male student, ninth grade, with a grade average of B+ in Civics. New Stanford Achievement Test social studies percentile is 88. This score is:

   a. Valid
   b. Questionable
   c. Invalid

5. A male student, sixth grade, with an average grade in language arts of C-. New CAT-77 reading vocabulary percentile is 24. This score is:

   a. Valid
   b. Questionable
   c. Invalid

6. A female student, ninth grade, with a D average in home economics. New IQ score is 129. This score is:

   a. Valid
   b. Questionable
   c. Invalid

7. A male student, first grade, has several notations of "needs improvement" in mathematics on his report card. New Metropolitan Achievement Test score in mathematics is 24th percentile. This score is:

   a. Valid
   b. Questionable
   c. Invalid

8. A female student, eighth grade, with an average grade of A. New CAT-77 reading comprehension percentile is 91. This score is:

   a. Valid
   b. Questionable
   c. Invalid

Appendix B

Pair Comparison Stimuli for Score Types

Name________________________________

Last four digits of
social security number________________

Directions

For each item, please select the answer you believe to be best, based on your own experience. There are no "right" or "wrong" answers. For each item, circle the letter of the answer you select.

There is a total of 10 items to be answered on 2 pages.

Please answer each item and do your own work.

Each year, the state sponsors testing of students in grades 4, 6, and 8 in basic skills on the California Achievement Test. Various types of scores are provided for students who take the test. For each of the following items, please select the type of score you believe would best help you, as an educator, to make sound decisions about what a student had or had not learned.

Please circle the letter of the type of score you select for each item.

Remember, you should choose the type of score YOU think would best help in making decisions about a student's skills.

1. a. Percentile rank (national)

   b. Scale score (ADSS--a C.A.T. scale)

2. a. The raw score (number right on test)

   b. Grade equivalent score

3. a. Grade-equivalent score

   b. Percentile rank

4. a. Scale score

   b. Stanine (national)

5. a. Stanine

   b. The raw score

6. a. Percentile rank

   b. Stanine

7. a. Grade-equivalent score

   b. Scale score

8. a. Stanine

   b. Percentile rank

9. a. Stanine

   b. Grade-equivalent score

10. a. Scale score

    b. The raw score

APPENDIX C

Pair Comparison Stimuli for Information Types

Name ______________________________

Last four digits of
social security number ________________

Directions

Read each item carefully and respond based on your own beliefs and experiences. There are no "right" or "wrong" answers.

There is a total of 21 items on 3 pages. For each item, circle the letter of the answer you select.

Please answer each item and do your own work.

When a new student comes to your school, some type of placement decision must be made. For each of the following questions, please circle the letter of the type of information you believe is likely to be MOST ACCURATE for making sound placement decisions.

1. a. The previous year's grades or marks.

   b. The previous year's standardized achievement test scores.

2. a. The results of an individual I.Q. test.

   b. The written recommendation of the last teacher.

3. a. The previous year's standardized achievement test scores.

   b. The parents' description of the child's school accomplishments.

4. a. The written recommendation of the previous school counselor.

   b. The results of an individual I.Q. test.

5. a. The previous year's local criterion-referenced achievement test scores.

   b. The results of an individual I.Q. test.

6. a. The parents' decription of the child's school accomplishments.

   b. The written recommendation of the last teacher.

7. a. The previous year's standardized achievement test scores.

   b. The written recommendation of the previous school counselor.

8. a. The previous year's grades or marks.

   b. The previous year's local criterion-referenced achievement test scores.

9. a. The written recommendation of the previous school counselor.

   b. The parents' description of the child's school accomplishments.

10. a. The parents' description of the child's school accomplishments.

b. The previous year's local criterion-referenced achievement test scores.

11. a. The results of an individual I.Q. test.

b. The previous year's grades or marks.

12. a. The written recommendation of the last teacher.

b. The written recommendation of the previous school counselor.

13. a. The previous year's standardized achievement test scores.

b. The results of an individual I.Q. test.

14. a. The written recommendation of the last teacher.

b. The previous year's local criterion-referenced achievement test scores.

15. a. The written recommendation of the previous school counselor.

b. The results of an individual I.Q. test.

16. a. The previous year's grades or marks.

b. The parents' description of the child's school accomplishments.

17. a. The previous year's local criterion-referenced achievement test scores.

b. The written recommendation of the previous school counselor.

18. a. The previous year's standardized achievement test scores.

b. The written recommendation of the last teacher.

19. a. The results of an individual I.Q. test.

b. The parents' description of the child's school accomplishments.

20. a. The written recommendation of the last teacher.

b. The previous year's grades or marks.

21. a. The previous year's local criterion-referenced achievement test scores.

b. The previous year's standardized achievement test scores.

APPENDIX D

Hypothetical Protocols

Hypothetical Protocol: Form A

Use the following information to answer items 1-2.

Student number 72-0013 Sex F Birth 04 / 16 / 70

Grade			Fall G.P.A.	Spring G.P.A.	Absent	Teacher
3	1978-79	Westside	82	80	9	Brooks
4	1979-80	Westside	79	83	8	Miller
5	1980-81	Westside	84			Smith

1979-80 MEAP CAT-77

	R-TOT	M-TOT	LA-TOT
ADSS	430	423	438
National Percentile	62	54	66
NCE	56	52	58

Otis-Lennon IQ/Spring 1979

89

MEAP SFTAA/Spring 1980

IQ	83
RSS	388

1. This information suggests that this student is performing at a level:

   a. Well above her ability
   b. About equal with her ability
   c. Well below her ability

2. Which type of information is likely to be the most reliable on this record?

   a. The G.P.A.
   b. The CAT-77 achievement subtest percentiles
   c. The Otis-Lennon IQ
   d. The SFTAA IQ

Hypothetical Protocol: Form B

Use the following information to answer items 1-2.

Student number 72-0013   Sex F   Birth 0 4 / 1 6 / 7 0

Grade			Fall G.P.A.	Spring G.P.A.	Absent	Teacher
3	1978-79	Westside	80	79	9	Brooks
4	1979-80	Westside	79	83	8	Miller
5	1980-81	Westside	78			Smith

1979-80 MEAP CAT-77

	R-TOT	M-TOT	LA-TOT
National Percentile	52	54	46
NCE	52	55	44

Otis-Lennon IQ/Spring 1979

110

MEAP SFTAA/Spring 1980

IQ	118
RSS	472

1. This information suggests that this student is performing at a level:

   a. Well above her ability
   b. About equal with her ability
   c. Well below her ability

2. Which type of information is likely to be the most reliable on this record?

   a. The G.P.A.
   b. The CAT-77 achievement test subtest percentiles
   c. The Otis-Lennon IQ
   d. The SFTAA IQ

Appendix E

Percentile Ranks Judgment Measure

Key to Percentile Ranks Judgment Measure

		Set			
Entry	z - score	Rank	±1/3 S.D. Band	±1/2 S.D. Band	±1 S.D. Band
1	-0.50	31	21-42	16-50	07-69
2	+2.00	98	96-99	93-99	84-99
3	+0.50	69	58-79	50-84	31-93
4	-1.50	07	04-12	02-16	01-31
5	0.00	50	36-62	31-69	16-84
6	+1.50	93	88-96	84-98	69-99
7	-1.00	16	10-24	07-31	02-50
8	-2.00	02	01-04	01-07	01-16
9	+1.00	84	76-90	69-93	50-98

Name: ____________________

Last Four Digits of
Social Security Number: ____

Directions

On the following sheets, you will find a number of test scores, expressed as percentile ranks or percentile bands.

Your percentile rank tells the percentage of a norm group that you have equaled or surpassed. For example, if your percentile rank for height in this class is 75, then you are as tall or taller than 75% of the persons in the class.

Because test scores tend to vary somewhat due to such chance factors as a lucky guess or the choice of questions, we sometimes express a score as a percentile band. The percentile band 50-75, for example, would mean that we are reasonably confident that the person earning this score is really better than the lower half of the group, but not as good as the top quarter of the group.

When the signal is given, open your booklet to page 1, and begin to work. Be sure that you finish each page before going on to the next page. DO NOT TURN BACK TO A PAGE ONCE YOU HAVE LEFT IT. WAIT FOR THE SIGNAL TO START.

Rating Key: 5=Score is well above mean
4=Score is somewhat above mean
3=Score is equal or nearly equal to mean
2=Score is somewhat below mean
1=Score is well below mean

Percentile Rank	Rating (Circle one for each given rank)				
31	1	2	3	4	5
98	1	2	3	4	5
69	1	2	3	4	5
07	1	2	3	4	5
50	1	2	3	4	5
93	1	2	3	4	5
16	1	2	3	4	5
02	1	2	3	4	5
84	1	2	3	4	5

Rating Key: 5=Score is well above mean
4=Score is somewhat above mean
3=Score is equal or nearly equal to mean
2=Score is somewhat below mean
1=Score is well below mean

Percentile Band	Rating (Circle one for each given band)				
21-42	1	2	3	4	5
96-99	1	2	3	4	5
58-79	1	2	3	4	5
04-12	1	2	3	4	5
38-62	1	2	3	4	5
88-96	1	2	3	4	5
10-24	1	2	3	4	5
01-04	1	2	3	4	5
76-90	1	2	3	4	5

Rating Key: 5=Score is well above mean
4=Score is somewhat above mean
3=Score is equal or nearly equal to mean
2=Score is somewhat below mean
1=Score is well below mean

Percentile Band	Rating (Circle one for each given band)				
16-50	1	2	3	4	5
93-99	1	2	3	4	5
50-84	1	2	3	4	5
02-16	1	2	3	4	5
31-69	1	2	3	4	5
84-98	1	2	3	4	5
07-31	1	2	3	4	5
01-07	1	2	3	4	5
69-93	1	2	3	4	5

Rating Key: 5=Score is well above mean
4=Score is somewhat above mean
3=Score is equal or nearly equal to mean
2=Score is somewhat below mean
1=Score is well below mean

Percentile Band	Rating (Circle one for each given band)				
07-69	1	2	3	4	5
84-99	1	2	3	4	5
31-93	1	2	3	4	5
01-31	1	2	3	4	5
16-84	1	2	3	4	5
69-99	1	2	3	4	5
02-50	1	2	3	4	5
01-16	1	2	3	4	5
50-98	1	2	3	4	5

Appendix F

Loss Ratio and Likelihood Ratio Estimate Measures

Key to Loss Ratio and
Likelihood Ratio Stimuli : Both Versions

		Responses	
Loss Ratio Item	Test Item	False Positive	False Negative
1	4	a	b
2	5	a	b
3	6	b	a
4	7	a	b
5	8	b	a
6	9	b	a

		Responses	
Likelihood Ratio Item	Test Item	False Positive	False Negative
1	10	a	b
2	11	b	a
3	12	b	a
4	13	a	b
5	14	a	b
6	15	b	a
7	16	a	b

FORM A: ATYPICAL STUDENT VERSION

Name ______________________

Last four digits of
social security number__________

Directions

Read each item carefully and respond based on your own beliefs and experiences. Except for the last four questions, there are no "right" or "wrong" answers.

For items 1-3, you will have to write in your response. For items 4-21, please circle the letter of the answer you select.

Please attempt every item and do your own work.

For questions 1-3, please choose your answers so that the numbers sum to 100%.

Based on your own experiences and observations, when students take the California Achievement Test:

1. What percent of these students receive a score which is a fairly accurate reflection of their skills? ____________%

2. What percent of these students receive a score which is much lower than their true capability? ____________%

3. What percent of these students receive a score which is much higher than their true capability? ____________%

TOTAL 100 %

For questions 4-9, select the statement which you believe is the WORSE of the pair of statements.

4. Which is WORSE:

   a. Accidentally placing a poor student in an advanced group or class.

   b. Accidentally placing a good student in a remedial group or class.

5. Which is WORSE:

   a. A student passing a test who really doesn't know the material.

   b. A student failing a test who really does know the material.

6. Which is WORSE:

   a. Making a student review material even though he or she knows the material well.

   b. Advancing a student to new material before he or she is ready.

7. Which is WORSE:

   a. Moving through material too quickly for most of the students.

   b. Moving through material too slowly for most of the students.

8. Which is WORSE:

a. A student just barely failing a test who probably knows the material.

b. A student just barely passing a test who probably does not know the material.

9. Which is WORSE:

a. A student forced to re-study material in a unit even though he or she really understands it.

b. A student who is confused over the material in a unit because he or she didn't master earlier units.

For questions 10-16, select the statement which you believe is the MORE LIKELY of the pair of statements to occur. C.A.T. means California Achievement Test.

10. Which is MORE LIKELY:

a. A generally poor student turns in a very good paper.

b. A generally good student turns in a very poor paper.

11. Which is MORE LIKELY:

a. A very good student receives a C.A.T. test score which is far too low.

b. A very poor student receives a C.A.T. test score which is far too high.

12. Which is MORE LIKELY:

a. A very poor student receives a C.A.T. test score which is far too low.

b. A very good student receives a C.A.T. test score which is far too high.

13. Which is MORE LIKELY:

a. A generally poor student performs very well on a classroom test.

b. A generally good student performs very poorly on a classroom test.

14. Which is MORE LIKELY:

a. A student who should be given a failing semester grade is passed.

b. A student who should be given a passing semester grade is failed.

15. Which is MORE LIKELY:

a. A student is placed in a group or class which is too low.

b. A student is placed in a group or class which is too high.

16. Which is MORE LIKELY:

a. A student who should fail a classroom test somehow passes.

b. A student who should pass a classroom test somehow fails.

17. When a new student comes to your school, what type of information is most likely to be most accurate for making a placement decision?

a. The previous year's grades or marks.

b. The previous year's standardized achievement test scores.

c. The written recommendation of the last teacher.

d. The parents' description of the child's accomplishments.

e. The results of an individual I.Q. test.

18. Have you ever taken a college or graduate course in Tests and Measurement?

a. Yes

b. No

19. What is the highest current certification which you hold?

a. A

b. AA

c. AAA

d. No current certification

20. Which best describes your school job?

a. Mostly or entirely teaching duties.

b. Mostly or entirely administrative duties.

c. About equally divided between teaching and administrative duties.

21. Would you like a summary of the results of this survey when it is complete?

a. Yes

b. No

FORM B: AVERAGE STUDENT VERSION

Name ______________________________

Last four digits of
social security number ____________

Directions

Read each item carefully and respond based on your own beliefs and experiences. Except for the last four questons, there are no "right" or "wrong" answers.

For items 1-3, you will have to write in your response. For items 4-20, please circle the letter of the answer you select.

Please attempt every item and do your own work.

For questions 1-3, please choose your answers so that the numbers sum to 100%.

Based on your own experiences and observations, when students take the California Achievement Test:

1. What percent of these students receive a score which is a fairly <u>accurate</u> reflection of their skills? __________%

2. What percent of these students receive a score which is much lower than their true capability? __________%

3. What percent of these students receive a score which is much higher than their true capability? __________%

TOTAL 100 %

For questions 4-9, select the statement which you believe is the WORSE of the pair of statements. For each question, the <u>student</u> is of AVERAGE achievement level.

4. Which is WORSE:

   a. Accidentally placing student in an advanced group or class.

   b. Accidentally placing student in a remedial group or class.

5. Which is WORSE:

   a. A student passing a test who really <u>doesn't</u> know the material.

   b. A student failing a test who really <u>does</u> know the material.

6. Which is WORSE:

   a. Making a student review material even though he or she knows the material well.

   b. Advancing a student to new material before he or she is ready.

7. Which is WORSE:

   a. Moving through material too quickly for most of the students.

   b. Moving through material too slowly for most of the students.

8. Which is WORSE:

   a. A student just barely failing a test who probably knows the material.

   b. A student just barely passing a test who probably does not know the material.

9. Which is WORSE:

   a. A student forced to re-study material in a unit even though he or she really understands it.

   b. A student who is confused over the material in a unit because he or she didn't master earlier units.

For questions 10-16, select the statement which you believe is the MORE LIKELY of the pair of statements to occur. C.A.T. means California Achievement Test. For each question, the student is of AVERAGE achievement level.

10. Which is MORE LIKELY:

    a. A student turns in a very good paper.

    b. A student turns in a very poor paper.

11. Which is MORE LIKELY:

    a. A student receives a C.A.T. test score which is far too low.

    b. A student receives a C.A.T. test score which is far too high.

12. Which is MORE LIKELY:

    a. A student receives a C.A.T. test score which is slightly low.

    b. A student receives a C.A.T. test score which is slightly high.

13. Which is MORE LIKELY:

    a. A student performs very well on a classroom test.

    b. A student performs very poorly on a classroom test.

14. Which is MORE LIKELY:

a. A student who should be given a failing semester grade is passed.

b. A student who should be given a passing semester grade is failed.

15. Which is MORE LIKELY:

a. A student is placed in a group or class which is too low.

b. A student is placed in a group or class which is too high.

16. Which is MORE LIKELY:

a. A student who should fail a classroom test somehow passes.

b. A student who should pass a classroom test somehow fails.

17. Have you ever taken a college or graduate course in Tests and Measurement?

a. Yes

b. No

18. What is the highest current certification which you hold?

a. A

b. AA

c. AAA

d. No current certification

19. Which best describes your school job?

a. Mostly or entirely teaching duties.

b. Mostly or entirely administrative duties.

c. About equally divided between teaching and administrative duties.

20. Would you like a summary of the results of this survey when it is complete?

a. Yes

b. No

Appendix G

Estimation of Pupil Performance Stimuli and Sample Booklet

Key to Estimation of
Pupil Performance Stimuli

Order	Stimulus	Initial Valence	Reliability	Follow-up Valence
1	Initial	Positive	N/A	N/A
2	Initial	Negative	N/A	N/A
3	Follow-up	N/A	Reliable	Positive
4	Follow-up	N/A	Reliable	Negative
5	Follow-up	N/A	Unreliable	Positive
6	Follow-up	N/A	Unreliable	Negative

Booklet is example of the following order: 1, 5 for a male student.

Stimuli

1. Carol is ten years old and beginning the fifth grade. She lives with her parents, an older brother, and two younger sisters. In an interview with her parents, her father gave his occupation as an engineer in an aerodynamics firm. In the interview her parents also noted that Carol spent about two hours each evening on her homework and reading books. On an individual intelligence test, Carol scored quite high.

   Note: All stimuli were generated for both a male student (Andrew) and a female student (Carol).

2. Carol is ten years old and beginning the fifth grade. She lives with her parents, an older brother, and two younger sisters. In an interview with her parents, her father gave his occupation as a machinist for an aerodynamics firm. In the interview, her parents also noted that Carol never did any homework but spent two hours each evening watching television. On an individual intelligence test, Carol scored quite low.

3. At mid-semester Andrew was tested in math and reading. The results showed that he was performing at about seventh grade level, approximately two years ahead of expectations for his age. The school psychologist reported that Andrew's curiosity enhanced his ability to do well in his math and reading, and that he had an enthusiastic and positive attitude toward school.

4. At mid-semester, Carol was tested in math and reading. The results showed that she was performing at about third grade level, approximately two years behind expectations for her age. The school psychologist reported that Carol had difficulty in directing her curiosity to school activities, often becoming distracted and losing interest in class discussions, and that she had a negative attitude toward school.

5. When interviewed, some of Carol's classmates said that they liked her and that they thought she was a good student. Cathy Robbins, an education student at a nearby college, had been hired as a substitute aid at Carol's school. She had assisted in Carol's class for a few days and had decided to administer an inkblot test to the class. She interpreted the results to mean that Carol was curious and enthusiastic about academic activities and that she had a positive attitude toward school.

6. When interviewed, some of Carol's classmates said that they didn't particularly like her and that they thought she wasn't a very good student. Cathy Robbins, an education student at a nearby college, had been hired as a substitute aid in Carol's school. She had assisted in Carol's class for a few days and decided to administer an inkblot test to the class. She interpreted the results to mean that Carol's curiosity led her to be easily distracted from academic activities and that she had a negative attitude toward school.

Sample Booklet: Estimation of Pupil Performance

Name ______________________________

Last four digits of
social security number __________

Directions

Attached is information concerning a new student. You are to read the information carefully, then answer the questions which follow. Please answer all questions with the response you believe to be best.

Once you have turned a page, do not turn back.

Begin when your instructor tells you to start.

Andrew is ten years old and beginning the fifth grade. He lives with his parents, an older brother, and two younger sisters. In an interview with his parents, his father gave his occupation as an engineer in an aerodynamics firm. In the interview his parents also noted that Andrew spent about two hours each evening on his homework and reading books. On an individual intelligence test, Andrew scored quite high.

Please turn to the next page and answer the questions.

Note: For items 2-4, circle the letter of the answer you choose.

1. What are the chances (between 0% and 100%) that Andrew will get mostly A's and B's on his report card?

   (Please write in your estimate) ____________________%

2. In selecting instructional materials for Andrew in reading and math at the beginning of the semester, what kinds of texts and instructional aids would you primarily use?

   a. Fifth grade level
   b. Fifth grade level and/or higher level
   c. Fifth grade level and/or lower level

3. Suppose that, during a math lesson, you asked Andrew a question and he hesitated: Would you:

   a. rephrase the same question in order to clarify it
   b. ask a similar question that is easier to answer
   c. further explain the problem, then repeat the same question
   d. ask the same question to another student
   e. answer the question yourself

4. How important is it for Andrew that you make a point of praising him every time he does good work?

   a. very important
   b. important
   c. somewhat important
   d. somewhat unimportant
   e. not important at all

When interviewed, some of Andrew's classmates said that they liked him and that they thought he was a good student. Cathy Robbins, an education student at a nearby college, had been hired as a substitute aid at Andrew's school. She had assisted in Andrew's class for a few days and had decided to administer an inkblot test to the class. She interpreted the results to mean that Andrew was curious and enthusiastic about academic activities and that he had a positive attitude toward school.

Please turn to the next page and answer the questions.

Note: For items 2-4, circle the letter of the answer you choose.

1. What are the chances (between 0% and 100%) that Andrew will get mostly A's and B's on his report card?

   (Please write in your estimate) ____________%

2. In selecting instructional materials for Andrew in reading and math at the beginning of the semester, what kinds of texts and instructional aids would you primarily use?

   a. Fifth grade level
   b. Fifth grade level and/or higher level
   c. Fifth grade level and/or lower level

3. Suppose that, during a math lesson, you asked Andrew a question and he hesitated. Would you:

   a. rephrase the same question in order to clarify it
   b. ask a similar question that is easier to answer
   c. further explain the problem, then repeat the same question
   d. ask the same question to another student
   e. answer the question yourself

4. How important is it for Andrew that you make a point of praising him every time he does good work?

   a. very important
   b. important
   c. somewhat important
   d. somewhat unimportant
   e. not important at all

Appendjx H

Test Knowledge and Perception Subscales

Key to Knowledge and Perception Subscales

Subscale	Number of items	Where found	Notes

Validity	8	Appendix A	Each of these items presented,nontest and test information. Responses were coded as:
			Valid = 3
			Questionable = 2
			Invalid = 1
			The combinations of information types, in order, were: LO-HI, HI-LO, HI-LO, HI-HI, LO-LO, LO-HI, HI-LO, HI-HI where the first entry is the nontest information and the second is the test score.
			Possible score range was, therefore, from 8-24, where high score represents belief in validity of test score (in conjunction with given nontest score).

Knowledge	10	P. 1-10	Each item was simply scored as right or wrong. The keyed responses were, in order: E, D, B, D, E, B, B, D, C, D

Key to Knowledge and Perception Subscales (continued)

Subscale	Number of items	Where found	Notes
Test-wiseness	14	II. 1-14	

Item	Key or Guide
1	Any response ("guess")
2	C (Absurd alternatives, plus clue or "ologist")
3	B (Stem asks for one meaning)
4	B (Answer is given by item 8)
5	A ("Smallest" number is the clue)
6	A (Answers B, C, D mean the same thing)
7	C (Correct response is very different in length)
8	B (Grammatical clue: "a")
9	B (Correct response is very different in length)
10	C (Resemblance of stem and correct alternative)
11	B (Stem asks for two outcomes, only 'B' gives two)
12	D (A, B, C mean the same thing)
13	B (Answer is given away by choices in item 1)
14	D (A, B, C mean the same thing)

Subscale	Number of items	Where found	Notes
Preference	5/9	III. A III. B	

The items used in this section were given weights to reflect the relative degree of dependence upon test score, as opposed to nontest score, information. The rating scale weights were as follows:

A. If the nontest information was used for the decision:
 1 = Incomplete nontest data, complete test data presented
 2 = Incomplete nontest data and incomplete test data presented
 3 = Complete nontest data and complete test data presented
 4 = Complete nontest data and incomplete test data presented

Key to Knowledge and Perception Subscales (continued)

Subscale	Number of items	Where found	Notes

B. If the test information was used for the decision:
5 = Complete test data and incomplete non-test data presented
6 = Complete test data and complete nontest data presented
7 = Incomplete test data and incomplete non-test data presented
8 = Incomplete test data and complete non-test data presented

The rating scale, from 1-8, represents increasingly higher degrees of dependence upon test data as the value goes up. Lower ratings represent a lower degree of dependence upon test data. The overall preference score was calculated as a percentage of the maximum possible score for each form of the instrument. High percentages would indicate strong dependence upon the test score data. Only those items for which conflicting information was presented were scored.

Part	Form	Item	Response Ratings
II	A	23	A = 7, B = 2
		24	A = 2, B = 7
		28	A = 1, B = 5
		29	A = 1, B = 5
		33	A = 8, B = 4

Part	Form	Item	Response Ratings
II	B	23	A = 6, B = 3
		24	A = 5, B = 1
		26	A = 8, B = 4
		27	A = 4, B = 8
		29	A = 8, B = 4
		31	A = 6, B = 3
		32	A = 6, B = 3
		33	A = 3, B = 6
		34	A = 4, B = 8

I. Knowledge

Use the following data to answer items 1-2.

Student	PRETEST ITEM					POSTTEST ITEM				
	1	2	3	4	5	1	2	3	4	5
John	+	+	0	0	0	+	+	+	0	+
Ann	0	+	+	0	0	+	0	+	0	0
Susan	+	+	+	0	0	+	+	+	+	+
Bill	0	0	0	0	0	+	+	+	0	+
Pete	+	0	0	+	0	+	+	+	+	+

+ = correct response; 0 = incorrect response

1. Which item shows greatest sensitivity to instruction?

   a. 1
   b. 2
   c. 3
   d. 4
   e. 5

2. If each item represents a different skill, what skill was learned (or taught) least well?

   a. 1
   b. 2
   c. 3
   d. 4
   e. 5

3. A particular test item has a difficulty index of .36. Teacher A says this means that 36% of the examinees missed the item. Teacher B says this item was a hard one for the examinees. Who is correct?

   a. A only
   b. B only
   c. both A and B
   d. neither A nor B

4. A student receives a percentile rank of 74 on a social studies achievement test. Teacher A says this means that 74% of the norms group did as well or better than this student. Teacher B says the student got 74% of the items correct. Who is correct?

   a. A only
   b. B only
   c. both A and B
   d. neither A nor B

5. On the CTBS (California Tests of Basic Skills), John obtains a raw score of 54 in mathematics concepts. On the CAT (California Achievement Test), Bill obtains a raw score of 44 in mathematics concepts. One appropriate conclusion is:

   a. John is more proficient in mathematics concepts than Bill.
   b. John and Bill are equally proficient in mathematics concepts.
   c. John's true score in mathematics concepts is higher than Bill's.
   d. John answered a larger proportion of items correctly than did Bill.
   e. No comparison should be made between these two scores.

6. A child performs at the 37th percentile on a nationally normed achievement test. If the child's ranking had been incorrectly determined by referring to a norms table for <u>schools</u>, the resulting percentile rank would be:

   a. higher
   b. lower
   c. unchanged

7. On an achievement test, two fourth grade students, Peter and Jane, received grade-equivalent scores of 4.4 and 8.2, respectively. Teacher A says Jane did as well on the test as the average eigth-grade students. Teacher B says Peter answered fewer items correctly than Jane. Who is correct?

   a. A only
   b. B only
   c. both A and B
   d. neither A nor B

8. A student received a grade-equivalent score of 10.2. This score indicates that:

   a. He ranks in his class at the equivalent of a rank of 10.2 for the grade 10 students of the normative group.
   b. He should be placed in the tenth grade in instruction in this subject.
   c. His raw score is the same as the median score earned by all students in the norm group who were 10.2 years old at the time of testing.
   d. His raw score on this test is the same as the approximate median of scores made by pupils in the second month of the tenth grade.

9. Which of the following indicates the BETTER performance on a normed test?

   a. A percentile rank of 65
   b. An NCE score of 40
   c. A T-score of 60
   d. There is no way to distinguish among the scores.

10. Which of the following indicates the POORER performance on a normed test?

   a. A 68% confidence band in percentiles of 38-54
   b. An NCE score of 45
   c. A 95% confidence band in percentiles of 30-62
   d. There is no way to distinguish among the scores.

II. Test-Wiseness

For items 1-14, choose the best answer. Each item except one suffers from a common item construction flaw.

1. One resistor of 30 ohms is wired in parallel with a resistor of 60 ohms. What is the total resistance?

   a. 20 ohms
   b. 45 ohms
   c. 60 ohms
   d. 90 ohms

2. An ornithologist is a person who

   a. sells shoes.
   b. drives a taxi cab.
   c. studies birds.
   d. plays a violin.

3. What is one meaning of the word panache?

a. ormolu or frantic
b. a bunch of feathers on a helmet
c. pandemonium or hoopla
d. helve, fractious, and chanteuse

4. Which of the following means "How are you?"

a. Maintenant, aujourd'hui?
b. Comment-allez vous?
c. Ne'est-ce pas?
d. Tres bien, et vous?

5. If you had one hour to answer fifty (50) multiple-choice questions, what is the smallest number you should have answered in a half-hour?

a. 10
b. 25
c. 30
d. 45
e. 50

6. When Bestor crystals are added to water,

a. the water turns blue.
b. the temperature rises.
c. heat is given off.
d. a thermometer will read higher.

7. How have scientists recognized the great work of Linnaeus?

a. By giving him the Nobel prize.
b. By founding a college with his name.
c. By adding the letter L. to the names of all the animals he had classified.
d. By awarding him a cash prize.

8. "Comment-allez vous?" which means "How are you?" is a:

a. old English saying.
b. French expression.
c. Italian phrase.
d. Arabic question.

9. To change a verb like cook to a gerund, you could

a. double the consonant and add the letters -ies.
b. add -ing.
c. change the verb to a predicate adjective, such as pressure cooker.
d. capitalize the first letter, and add -ed to the word or sentence.

10. Another word for convivial is

a. voracious.
b. inextricable.
c. jovial.
d. placebo.

11. In the southern United States, two outcomes of the Civil War were

a. slavery flourished in most states.
b. reconstruction and the abolition of slavery.
c. more wars in mainland China during 1871-1880.
d. fewer plantations in Alabama.

12. If something is flammable, it will

a. resist burning.
b. not catch on fire.
c. not be consumed by flames.
d. easily ignite.

13. If a resistor of 60 ohms is wired in parallel with a resistor of 30 ohms, the total resistance is 75 ohms (60 + ½ x 30).

a. True
b. False

14. An example of an opening in a room is

a. a window.
b. an egress.
c. a doorway.
d. all of the above.

III. A. Preference

For items 21-35, you are to read each record card for a student. All scores given are percentile ranks. The interest area scores are from the Kuder Preference Record, Vocational Form C.

The record also gives evaluations of the student by the adviser. The advisers all have considerable teaching experience as well as training in educational and vocational guidance.

The names are fictitious, but otherwise the records are accurate. All data on the records were obtained during the tenth grade year.

You are to decide whether the student should be placed in the regular or accelerated science class for grade 11. In the accelerated class, students are expected to learn at a faster rate and more intensively than in the regular class.

You should examine the information for each student, then decide for which class you will recommend the student. Mark that choice on your answer sheet.

There is no limit on the number of students you place in either science class.

21. YEARLY RECORD FOR GRADE 10

NAME Gregory Barton AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
		Reading	65	Mechanical	75
		Science	89	Computational	87
		Math	88	Scientific	83
		Social Studies	64	Persuasive	64
				Artistic	50
				Literary	43
				Musical	36
				Social Service	28
				Clerical	19

HOME-ROOM TEACHER: An excellent student high in achievement and ability.

ADVISER: Well-liked. Capable. Conscientious. Excellent student.

a. Accelerated science
b. Regular science

22. YEARLY RECORD FOR GRADE 10

NAME Glen Chapman AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	109	Reading	26	Mechanical	23
		Science	25	Computational	21
		Math	26	Scientific	18
		Social Studies	24	Persuasive	41
				Artistic	63
				Literary	61
				Musical	48
				Social Service	83
				Clerical	87

HOME-ROOM TEACHER: Glen has his heart set on becoming a scientist like his father. Unfortunately his ability does not seem to warrant this. He accompanies his father to the lab evenings and weekends and loves every minute of it. He works very hard but does not seem to understand basic scientific concepts.

ADVISER: Glen is keenly interested in all things scientific. All three science teachers have commented to me on his interest but they are worried that his ability is just not up to his ambitions.

a. Accelerated science
b. Regular science

23. YEARLY RECORD FOR GRADE 10

NAME Doris Shechan AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	124	Reading	84		
		Science	82		
		Math	81		
		Social Studies	84		

HOME-ROOM TEACHER: This girl has no interest in anything but athletics. She spends all of her time in the gym. Her English teacher tells me she writes nearly all of her papers on games and sports.

ADVISER: Interested only in sports. I have talked with her about becoming a physical education teacher but she says she wants to "play," not "teach."

a. Accelerated science
b. Regular science

24. YEARLY RECORD FOR GRADE 10

NAME John Dewitt AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	106	Reading	39		
		Science	26		
		Math	27		
		Social Studies	44		

HOME-ROOM TEACHER: John cares only for science. He is never happier than when he is "experimenting" in the little laboratory he built in his basement at home.

ADVISER: Very interested in science. He told me that his chief problem was to decide which field of science to go into.

a. Accelerated science
b. Regular science

25. YEARLY RECORD FOR GRADE 10

NAME Mary Mullen AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	129			Mechanical	26
				Computational	29
				Scientific	32
				Persuasive	43
				Artistic	76
				Literary	54
				Musical	40
				Social Service	65
				Clerical	94

HOME-ROOM TEACHER: Every teacher who has this girl complains about her. She is near the bottom in all her classes; her work is rarely handed in on time; she practically refuses to recite or to answer when called on.

ADVISER: I am concerned about Mary. She has no interest, no plans, no ambitions. She dislikes school intensely and refuses to work at anything. A very difficult girl.

a. Accelerated science
b. Regular science

26. YEARLY RECORD FOR GRADE 10

NAME Elaine Humphrey AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	120	Reading	81	Mechanical	85
		Science	82	Computational	87
		Math	82	Scientific	85
		Social Studies	84	Persuasive	16
				Artistic	49
				Literary	37
				Musical	43
				Social Service	21
				Clerical	31

HOME-ROOM TEACHER:

ADVISER:

a. Accelerated science
b. Regular science

27. YEARLY RECORD FOR GRADE 10

NAME Margaret Hilton AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	103			Mechanical	82
				Computational	81
				Scientific	79
				Persuasive	58
				Artistic	42
				Literary	46
				Musical	48
				Social Service	60
				Clerical	22

HOME-ROOM TEACHER: Excellent student. The math teacher tells me that he has yet to call on Margaret for an explanation that she cannot provide.

ADVISER: A born mathematician. Bright and capable girl. Will do well in any type of scientific research.

a. Accelerated science
b. Regular science

28. YEARLY RECORD FOR GRADE 10

NAME Margaret Nielson AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	109	Reading	23	Mechanical	21
		Science	26	Computational	17
		Math	24	Scientific	26
		Social Studies	26	Persuasive	40
				Artistic	37
				Literary	59
				Musical	63
				Social Service	25
				Clerical	87

HOME-ROOM TEACHER: Margaret is a capable and industrious student. She does good work in all her classes and is very popular with both her teachers and her peers.

ADVISER: This girl has yet to make a firm decision regarding her future. Her chief interest lies in working in a hospital, but she does not want to become a nurse. I have discussed the possibilities of her becoming a laboratory technician, an X-ray technician, or doing medical research. Of these she prefers the last. Her interest and capability in science would make this a good choice for her.

a. Accelerated science
b. Regular science

29. YEARLY RECORD FOR GRADE 10

NAME Mildred Learch AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	106	Reading	23	Mechanical	67
		Science	25	Computational	81
		Math	23	Scientific	93
		Social Studies	26	Persuasive	63
				Artistic	39
				Literary	41
				Musical	16
				Social Service	32
				Clerical	19

HOME-ROOM TEACHER: A superior student. Does excellent work in all of her classes.

ADVISER: One of our better students. No definite plans other than "college" as yet.

a. Accelerated science
b. Regular science

30. YEARLY RECORD FOR GRADE 10

NAME Ruth Skillman AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	106	Reading	73	Mechanical	88
		Science	85	Computational	81
		Math	88	Scientific	84
		Social Studies	76	Persuasive	48
				Artistic	53
				Literary	41
				Musical	37
				Social Service	47
				Clerical	55

HOME-ROOM TEACHER: This girl's ability is quite high. On two different occasions, teachers have told me that when class discussion gets involved she can ask a question that cuts right to the heart of the matter.

ADVISER: This girl wants to become a high-school teacher and I have encouraged her in this. She is of superior ability and I believe she will be quite successful in working with students.

a. Accelerated science
b. Regular science

31. YEARLY RECORD FOR GRADE 10

NAME Morton Dawson AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
				Mechanical	17
				Computational	28
				Scientific	31
				Persuasive	62
				Artistic	24
				Literary	23
				Musical	19
				Social Service	48
				Clerical	71

HOME-ROOM TEACHER: Poor student. Limited ability.

ADVISER: Plans to become a chemist like his father and brother but his low ability and achievement make this possibility unlikely.

a. Accelerated science
b. Regular science

32. YEARLY RECORD FOR GRADE 10

NAME Catherine Kenny AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
		Reading	26	Mechanical	23
		Science	26	Computational	21
		Math	24	Scientific	18
		Social Studies	23	Persuasive	49
				Artistic	53
				Literary	57
				Musical	36
				Social Service	72
				Clerical	89

HOME-ROOM TEACHER: Catherine is a very conscientious student who gets along well with everyone. Although she works very hard and gets good marks she does not always seem to "grasp" the essentials.

ADVISER: Is seriously considering becoming a high-school science teacher.

a. Accelerated science
b. Regular science

33. YEARLY RECORD FOR GRADE 10

NAME Martin Anderson AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	121				

HOME-ROOM TEACHER: This boy is near the bottom of his class in achievement. Many teachers have commented to me about his poor work.

ADVISER: Poor worker. Very low in achievement. Interested only in athletics. Talks of being a professional athlete.

a. Accelerated science
b. Regular science

34. YEARLY RECORD FOR GRADE 10

NAME Burt Ingram AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank

HOME-ROOM TEACHER: Inferior ability and achievement. Does failing work in most of his classes.

ADVISER: No interest in school or any of his classes. Spends most of his time with his gang hanging around street corners. Below average in ability and achievement.

a. Accelerated science
b. Regular science

35. YEARLY RECORD FOR GRADE 10

NAME Bill Turner AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of		Reading	64	Mechanical	36
Mental Maturity:	123	Science	44	Computational	50
		Math	41	Scientific	41
		Social Studies	72	Persuasive	63
				Artistic	81
				Literary	78
				Musical	82
				Social Service	46
				Clerical	79

HOME-ROOM TEACHER: Bill's ability is questionable. His teachers tell me that they frequently doubt that the work he hands in is his own. He rarely recites in class or enters into the discussion, and when called on he seems not to understand the question.

ADVISER: Bill's parents have talked with me about whether to send him to college, but I doubt that he has the ability. Various comments about his behavior in class from his teachers tend to support my judgment in this.

a. Accelerated science
b. Regular science

III. B. Preference

For items 21-35, you are to read each record card for a student. All scores given are percentile ranks. The interest area scores are from tne Kuder Preference Record, Vocational Form C.

The record also gives evaluations of student by the adviser. The advisers all have considerable teaching experience as well as training in educational and vocational guidance.

The names are fictitious, but otherwise the records are accurate. All data on the records were obtained during the tenth grade year.

You are to decide whether the student should be placed in the regular or accelerated science class for grade 11. In the accelerated class, students are expected to learn at a faster rate and more intensively than in the regular class.

You should examine the information for each student, then decide for which class you will recommend the student. Mark that choice on your answer sheet.

There is no limit on the number of students you place in either science class.

21. YEARLY RECORD FOR GRADE 10

NAME Gregory Barton AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
		Reading	65	Mechanical	75
		Science	89	Computational	87
		Math	88	Scientific	83
		Social Studies	64	Persuasive	64
				Artistic	50
				Literary	43
				Musical	36
				Social Service	28
				Clerical	19

HOME-ROOM TEACHER: An excellent student high in achievement and ability.

ADVISER: Well-liked. Capable. Conscientious. Excellent student.

a. Accelerated science
b. Regular science

22. YEARLY RECORD FOR GRADE 10

NAME Glen Chapman AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity	109	Reading	26	Mechanical	23
		Science	25	Computational	21
		Math	26	Scientific	18
		Social Studies	24	Persuasive	41
				Artistic	63
				Literary	61
				Musical	48
				Social Service	83
				Clerical	87

HOME-ROOM TEACHER: Glen has his heart set on becoming a scientist like his father. Unfortunately his ability does not seem to warrant this. He accompanies his father to the lab evenings and weekends and loves every minute of it. He works very hard but does not seem to understand basic scientific concepts.

ADVISER: Glen is keenly interested in all things scientific. All three science teachers have commented to me on his interest but they are worried that his ability is just not up to his ambitions.

a. Accelerated science
b. Regular science

23. YEARLY RECORD FOR GRADE 10

NAME Paul Kilgore AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	121	Reading	81	Mechanical	85
		Science	83	Computational	87
		Math	84	Scientific	85
		Social Studies	82	Persuasive	41
				Artistic	16
				Literary	22
				Musical	19
				Social Service	38
				Clerical	21

HOME-ROOM TEACHER: I have heard two different teachers comment on Paul's lackadaisical attitude and class work, and I agree with them. His ability and achievement are both below average and his interest in his studies is nil.

ADVISER: Paul is a difficult boy to talk to. When I try to get at the reason for his poor school work and total lack of interest he clams up and I get no where. His lack of ability is as apparent to all of his teachers as it is to me.

a. Accelerated science
b. Regular science

24. YEARLY RECORD FOR GRADE 10

NAME Keith Warren AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	123	Reading	84	Mecnanical	33
		Science	81	Computational	45
		Math	81	Scientific	37
		Social Studies	83	Persuasive	81
				Artistic	69
				Literary	67
				Musical	52
				Social Service	86
				Clerical	49

HOME-ROOM TEACHER: This boy is extremely negative toward his work. He has come into serious conflict with two of his teachers. His achievement is very low, and in ability he is near the bottom of his class.

ADVISER:

a. Accelerated science
b. Regular science

25. YEARLY RECORD FOR GRADE 10

NAME Kathy Parker AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
				Mechanical	81
				Computational	79
				Scientific	84
				Persuasive	31
				Artistic	[illegible]
				Literary	79
				Musical	42
				Social Service	37
				Clerical	61

HOME-ROOM TEACHER: An excellent student. Stands high in all of her classes, but is especially interested in English and literature.

ADVISER: Plans to become a writer. Superior in ability and achievement. I have discussed colleges and college courses with her in detail.

a. Accelerated science
b. Regular science

26. YEARLY RECORD FOR GRADE 10

NAME Ruth Changer AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
		Reading	65	Mechanical	74
		Science	83	Computational	82
		Math	80	Scientific	86
		Social Studies	63	Persuasive	31
				Artistic	16
				Literary	25
				Musical	33
				Social Service	45
				Clerical	59

HOME-ROOM TEACHER: A bright girl but is below average in achievement. More interested in her duties as cheer-leader than in her school work.

ADVISER: A pleasant and popular girl. Does not work up to her full capability. Plans to become a beautician and work in her sister's beauty parlor.

a. Accelerated science
b. Regular science

27. YEARLY RECORD FOR GRADE 10

NAME Joyce Durwith AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Intarest Area	Percentile Rank
California Test of Mental Maturity:	109				

HOME-ROOM TEACHER: A very capable girl. Does well in all of her classes.

ADVISER: Very good student. Have talked with her about going on to college. She plans to study nuclear physics.

a. Accelerated science
b. Regular science

28. YEARLY RECORD FOR GRADE 10

NAME Alex Crane AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank

HOME-ROOM TEACHER: A top-notch student. Several teachers have commented to me about what a pleasure it is to have Alex in their classes. His work is always well done and always in on time. He seems interested in everything.

ADVISER: This boy's only problem is in deciding what most interests him. He enjoys all of his classes and does very good work in all of them. To date he has considered Law, Medicine, Politics, and Teaching!

a. Accelerated science
b. Regular science

29. YEARLY RECORD FOR GRADE 10

NAME Frances Delong AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	129	Reading	84		
		Science	82		
		Math	81		
		Social Studies	81		

HOME-ROOM TEACHER: This girl is a problem! Her work is very poor, her ability is definitely below average, and her attitude toward school and her teachers worse than both. Every teacher complains of her poor attitude and lack of interest.

ADVISER: If this girl has any interests I cannot locate them. I have talked with her several times, but no success. Her lack of ability and achievement are all part of the same picture.

a. Accelerated science
b. Regular science

30. YEARLY RECORD FOR GRADE 10

NAME Darrell O'Rourke AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
		Reading	28	Mechanical	42
		Science	17	Computational	39
		Math	14	Scientific	43
		Social Studies	26	Persuasive	84
				Artistic	68
				Literary	42
				Musical	27
				Social Service	65
				Clerical	79

HOME-ROOM TEACHER: Below average student, quite limited in ability and achievement. Careless about his work. Dislikes school.

ADVISER: Ability and achievement are both limited.

a. Accelerated science
b. Regular science

31. YEARLY RECORD FOR GRADE 10

NAME Bernice Eager AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
Kuhlmann-Anderson	122	Reading	84	Mechanical	87
		Science	84	Computational	85
		Math	84	Scientific	93
		Social Studies	81	Persuasive	40
				Artistic	27
				Literary	36
				Musical	31
				Social Service	43
				Clerical	22

HOME-ROOM TEACHER: Bernice is extremely bright. She loves her work in home economics and dreams of the day when she will have her own home and family. She has no interest in anything except home-planning and home-management.

ADVISER: This girl's strong interest in home economics and her very high ability has led me to suggest that she enter this field professionally. She will have none of it. She has no interest in anything other than becoming a wife and mother.

32. YEARLY RECORD FOR GRADE 10

NAME Carroll Scott AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	120	Reading	81	Mechanical	87
		Science	84	Computational	86
		Math	83	Scientific	93
		Social Studies	81	Persuasive	40
				Artistic	18
				Literary	26
				Musical	38
				Social Service	54
				Clerical	19

HOME-ROOM TEACHER: Below average in achievement. Work is sloppy and never on time. The only teacher who has not commented on this is the physical education teacher. She always gets A's in physical education.

ADVISER: This girl's low achievement will prevent her from being successful in college. She is planning to attend college, and I have several times warned her that unless her achievement improves she will have difficulty in gaining admittance. She plans to become a physical-education teacher.

a. Accelerated science
b. Regular science

33. YEARLY RECORD FOR GRADE 10

NAME Michael Vaughan AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
OTIS:	107	Reading	23	Mechanical	21
		Science	24	Computational	18
		Math	26	Scientific	24
		Social Studies	24	Persuasive	36
				Artistic	41
				Literary	32
				Musical	58
				Social Service	85
				Clerical	79

HOME-ROOM TEACHER: A very hard-working student. Gets good grades.

ADVISER: Mike plans to become a high-school science teacher and I have encouraged him in this. I talked with his chemistry teacher who told me of the excellent work Mike did on his science projects. It seems as though he spent more time and did a more thorough job than anyone else in the class.

a. Accelerated science
b. Regular science

34. YEARLY RECORD FOR GRADE 10

NAME Robert Elliott AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity:	108	Reading	24		
		Science	26		
		Math	26		
		Social Studies	23		

HOME-ROOM TEACHER: Robert is a capable and hard-working student. He does good work in all of his classes. His ability is well above average.

ADVISER: Plans to become a chemist or a physician. Does excellent work in his science classes.

a. Accelerated science
b. Regular science

35. YEARLY RECORD FOR GRADE 10

NAME Norman Richardson AGE: 16

Intelligence Test	IQ	Achievement Test	Percentile Rank	Kuder Interest Area	Percentile Rank
California Test of Mental Maturity	108	Reading	23	Mechanical	21
		Science	26	Computational	24
		Math	26	Scientific	24
		Social Studies	24	Persuasive	62
				Artistic	37
				Literary	39
				Musical	51
				Social Service	78
				Clerical	61

HOME-ROOM TEACHER:

ADVISER:

a. Accelerated science
b. Regular science