DOCUMENT RESUME

ED 209 344 TM 810 897

AUTHOR Cummings, Oliver W.
TITLE Validation of a Diagnostic Interpretation Technique for the Iowa Tests of Basic Skills: Final Report to the National Institute of Education.
INSTITUTION Grant Wood Area Education Agency. Cedar Rapids, Iowa.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
PUB DATE 30 Jun 81
GRANT NIE-G-80-0084
NOTE 71p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Rating; Achievement Tests; *Basic Skills; Elementary School Mathematics; Elementary School Students; *Intermediate Grades; *Scores; Student Evaluation; *Test Interpretation; *Test Results
IDENTIFIERS *Iowa Tests of Basic Skills; Stanford Diagnostic Mathematics Test

ABSTRACT

Three studies related to group interpretation of the subskills tested by the Iowa Tests of Basic Skills (ITBS) were the focus of this project. These were the Impact Study, Teachers' Predictions of Student Performance on Subskills of Mathematics, and Relationships between the Results of ITBS. Mathematics Subtests and Stanford Diagnostic Mathematics Test. Large groups of students were shown how to interpret their results from the ITBS through the use of the Pupil Item Response Record and a Skill Summary Sheet. A need exists for interpretation of results of school wide testing programs. Students who have gone through this interpretation process emerge with an improved self image. As discrepancies occur between teacher expectations and actual performance, they can be explained. This benefits both teachers and students. The interpretation process was designed to actively involve students in assessing their performance on a standardized achievement test. The student thereby becomes an active, rather than passive, recipient of test results. (DWH)

VALIDATION OF A DIAGNOSTIC INTERPRETATION TECHNIQUE FOR THE IOWA TESTS OF BASIC SKILLS: FINAL REPORT TO THE NATIONAL INSTITUTE OF EDUCATION

Oliver W. Cummings
Grant Wood Area Education Agency
Cedar Rapids, Iowa 52404

This research was supported, in part, by the National Institute of Education under Grant Number NIE-G-80-0084. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the author and do not necessarily reflect the views of the Institute or the Department of Education.

ACKNOWLEDGMENTS

I would like to use this opportunity to formally express my appreciation to the persons and organizations that made this research possible. Without the concern for better use of test data in the schools of eastern Iowa, and the active participation of the teachers in the seven school districts involved in the various studies, the research would not have been completed. Also, without the monetary support of the National Institute of Education, the research would not have been done.

In respect to people besides the teachers and students who gave of their time, I'd like to recognize Mrs. Marilyn Stoecker, of the Cedar Rapids Community School District, whose involvement in test interpretation for students provided the catalyst for the development of the interpretation technique; Mrs. P. K. Sommers, of the Grant Wood Area Education Agency, who provided particularly valuable assistance in doing the interpretation sessions and in data collection; Dr. Tom Stinard, of the Grant Wood Area Education Agency, who did much of the data analysis and assisted in the preparation of the final report; and Dr. H. D. Hoover, of the University of Iowa, who did some of the data analysis and provided valuable reviews of several parts of the report. Finally, I express my appreciation to Mrs. Mary Jo Sires, of the Grant Wood Area Education Agency, who very competently kept track of the forms, data, and other materials used throughout the year, and who patiently worked through the preparation of this manuscript.

While I owe these people a debt of gratitude for their participation, I am alone responsible for any errors or omissions in this document.

Oliver W. Cummings
June 30, 1981

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Validation of a Diagnostic Interpretation Technique
for The Iowa Tests of Basic Skills

Building upon the body of literature recommending the diagnostic use and interpretation of standardized achievement tests, this project focused on three studies related to group interpretation of the sub-skills tested by the Iowa Tests of Basic Skills (Hieronymus, Lindquist, & Hoover, 1978). The interpretation technique employed emphasized providing feedback to students and teachers about performance on each of 60 to 70 skills tested at the various levels of the standardized achievement test.

Study 1 was a study of the impact of the interpretation sessions on teachers and students. Both attitudes toward standardized tests and knowledge about the Iowa Tests of Basic Skills were assessed for experimental and control groups of teachers and students. Study 2 addressed the commonly held belief that teachers have fairly accurate perceptions of "how well students are doing" in their skills development. This study also explored the differences in estimations made under raw-score and norm-referenced frameworks, and on the effects of grade level, student sex, and overall mathematics achievement on the predictions. Study 3 was essentially a concurrent validity study between the Iowa Tests of Basic Skills and the Stanford Diagnostic Mathematics Test. Since the interpretation technique employed constituted a diagnostic use of the survey test, this study was included to address the question of whether it measured the same things in the same ways as the diagnostic test.

The findings of the three studies incorporated into this project led to a conclusion that there is a need for the interpretation of the results of tests administered in school-wide testing programs, and there was modest support for providing "diagnostic" interpretation of the "survey" test. At least two bases for this conclusion were found. First, students who have been through the interpretation process used, felt that they had done better on the test than students who had not had the test results interpreted to them. Secondly, the act of interpretation should raise important questions for the teachers, as discrepancies between expectations and actual performance occur. This should benefit both students and teachers as reasons for the discrepancies between the students' behaviors and the teachers' expectations are explained. The benefit for teachers should be an opportunity to: 1) reassess their expectations for certain students; and 2) examine some of their biases about the performance of certain subgroups in the subject areas tested. The benefit for students should be a better educational process borne out of higher expectations for themselves and more appropriate expectations from their teachers, regardless of the student's sex or overall achievement level.

VALIDATION OF A DIAGNOSTIC INTERPRETATION TECHNIQUE FOR THE IOWA TESTS OF BASIC SKILLS

PART I: INTRODUCTION

In 1972, the American Personnel and Guidance Association and the National Council on Measurement in Education adopted a joint resolution on the responsible use of tests. In part, their position statement reads:

> In schools and colleges the principal needs served by testing include the providing of information (1) to teachers as an aid to the improvement of instruction; (2) to students and, in the case of younger students, to their parents, as an aid to self-understanding and to both educational and vocational planning; and (3) to administrators, as a basis for planning, decision-making, and evaluating the effectiveness of programs and operations. (American Personnel and Guidance Association & National Council on Measurement in Education, 1972.)

Further, in 1980, the American Personnel and Guidance Association issued a policy statement titled "Responsibilities of Users of Standardized Tests" (American Personnel and Guidance Association, 1980). This policy statement emphasized the importance of the test user becoming familiar with the test, and the need for presentation of the test data so that it is comprehensible to the user.

The importance of comprehensible feedback to the person tested has been previously recognized as a condition of the ethical use of tests (Lyman, 1974), and as an important aspect of student motivation (Kirkland, 1971). Feedback is further recognized as one means of meeting test consumers' needs (Bradley, 1981), and as the core of Bradley's (1978) person-referenced test interpretation. Through person-referenced interpretation, items from a test are reviewed by the student and a counselor in an attempt to personalize the results for the student and to assist the student in processing the information gleaned through the interpretation. Bradley (1978) contends that it

is difficult to personalize performance and fully promote self-understanding using normative scores alone. Buros (1977), in the same vein, suggested that the recording of normed scores alone be replaced by compound scores consisting of the normed scores, the percentage of items estimated to be known, and, possibly the obtained percentile rank. The intent of the compound score is according to Buros, "to shift our emphasis from differentiation to measurement."

When moves away from the use of normative scores are made, there is a logical, following use of the test results for diagnostic purposes. At the heart of person-referenced test interpretation is an intention to allow the student to analyze correct and incorrect responses to individual items on the test and to respond to that analysis in a personal way. Presumably, a similar outcome would result if the focus of testing moved more toward measurement than differentiation. Diagnostic interpretation of tests has been encouraged by test publishers, through various raw-score and item-response report forms for their standardized, survey instruments. The claims made for these report forms range from providing clues for selective follow-up (Hieronymus, Lindquist, & Hoover, 1979, p. 31), to helping point out a student's relative strengths and weaknesses within a specific skill domain (Prescott, Balow, Hogan, & Farr, 1978, p. 33), to determining an individual's strengths and weaknesses in the various categories of skills tested (CTB/McGraw Hill, 1977, p. 65).

In general, the interpretation of item data and/or skill clusters for diagnostic uses has been widely recommended. Ebel (1972) suggests that "many achievement tests can provide 'diagnostic' information of value to the individual pupil if he is told which items he missed" (p. 478), and Rudman (1977) indicates that through scoring options available from test publishers, "classically constructed standardized achievement tests can be used analytically, for they can be referenced in one of several modes: by norms, by criteria, by objectives" (p. 181).

Building upon the body of literature recommending the diagnostic use and interpretation of standardized achievement tests, this project focused on three studies related to group interpretation of the sub-skills tested by the Iowa Tests of Basic Skills (Hieronymus, Lindquist, & Hoover, 1978). The interpretation technique employed emphasized

providing feedback to students and teachers about performance on each of 60 to 70 skills tested at the various levels of the standardized achievement test. This interpretation technique is described in Part II: The Interpretation Technique.

Study 1, which is described in Part III: The Impact Study, was a study of the impact of the interpretation sessions on teachers and students. Both attitudes toward standardized tests and knowledge about the Iowa Tests of Basic Skills were assessed for experimental and control groups of teachers and students.

Study 2, covered in Part IV: Teachers' Predictions of Student Performance on Subskills of Mathematics, addressed the commonly held belief that teachers have fairly accurate perceptions of "how well students are doing" in their skills development. This study also explored the differences in estimations made under raw-score and norm-referenced frameworks, and on the effects of grade level, student sex, and overall mathematics achievement on the predictions.

Study 3, presented in Part V: Relationships Between the Results of the Iowa Tests of Basic Skills, Mathematics Subtests, and the Stanford Diagnostic Mathematics Test, was essentially a concurrent validity study. Since the interpretation technique employed constituted a diagnostic use of the survey test, this study was included to address the question of whether it measured the same things in the same ways as the diagnostic test.

Finally, in Part VI: Summary of the Project and Conclusions, a discussion of the validity and usefulness of the interpretation technique is provided. In this discussion the pertinent findings of the three studies are integrated to highlight the uses of the technique and to point out the weaknesses and some cautions that should be observed. These considerations indicate some specific needs for further research on the process and include speculation about some aspects of the interpretation process which were not studied.

PART II: THE INTERPRETATION TECHNIQUE

The approach described here was conducted in classroom-sized groups, and required approximately 40 to 50 minutes per group. It was a time-efficient way to provide feedback to large numbers of

students, thus, meeting the professional responsibility of interpretation and freeing valuable teacher or counselor time for dealing with students who need extra help in understanding the test results. The interpretation process involved several steps, leading students to a summary of their own performance on each of approximately 60 skills identified on the Iowa Tests of Basic Skills. The students used their own scoring service report form (the Pupil Item Response Record) and a Skill Summary Sheet, which was constructed for the project.

Description of the Pupil Item Response Record

The Pupil Item Response Record provides complete information on each pupil's answer for each question on the test. Identifying information about the student, grade-equivalent (or other developmental) scores, and percentile ranks are given on the report. In addition, the percentage of correct and incorrect responses for each subtest, plus the item number, the student's response to each item (correct, incorrect, or omit), the difficulty level of the item, and the skill measured by the item are provided. Figure 1 shows a sample segment of the report for one of the eleven subtests on the Iowa Tests of Basic Skills main battery. A "+" indicates a correct answer, a "-" indicates an incorrect response, and an "0" represents a question left unanswered. The item numbers are read vertically and are out of sequence, because the scoring program clusters items for the same skill together. The difficulty scale runs from 1 to 9, with 1 being the most difficult (10-19 percent of students in the standardization sample selecting a correct response).

The skill codes are as follows:

1 = Single-step problems: addition or subtraction.

2 = Single-step problems: multiplication or division.

3 = Multiple-step problems: combined use of basic operations.

The secondary skill codes, C, W, and F, represent currency, whole numbers, and fractions, respectively. The secondary codes are not used in the technique presented here.

Figure 1. Sample Section of a Pupil Item Response Record for Level 11, Form 8 of the Iowa Tests of Basic Skills.

Test M-2 Mathematics Problems

GE = 52 PR = 50 % Correct = 52 % incorrect = 37 % NA = 11

item number	17	18	21	23	29	30	32	37	20	22	25	33	34	39	41	19	38	40	24	26	27	28	31	35	36	42	43
response	+	+	+	+	+	+	+	-	+	+	+	-	+	+	0	0	-	-	-	+	0	+	-	-	-	-	-
difficulty	7	7	7	6	7	7	6	6	8	5	5	5	5	5	3	8	3	3	8	6	4	6	4	5	6	5	5
skill	1 C	1 C	1 C	1 C	1 W	1 W	1 F	1 F	2 C	2 C	2 C	2 W	2 W	2 W	2 F	3 C	3 C	3 C	3 W	3 W	3 W	3 W	3 W	3 W	3 W	3 W	3 W

Description of the Skill Summary Sheet

The Skill Summary Sheet is a listing of the major skills tested on the Iowa Tests of Basic Skills. For each major skill tested, three broad categories of performance were defined: 1) Satisfactory progress likely; 2) More information needed; and 3) Possible problem area. The three categories were divided on the sheet according to raw score performance, which was adjusted for differences in the mean difficulty of the set of items used to test each skill.

In determining the raw score ranges for the three categories of performance, the "Satisfactory progress likely" column included raw scores generally at or above the mean raw score performance of the students in the norm group. The division between the remaining two categories was determined by allocating the items on a percentage basis of approximately half the difference between the mean percent correct and zero.

For example, a skill tested with 10 items and having a mean difficulty of 50 percent would have raw scores of 5 through 10 in the "Satisfactory progress likely" column, scores of 2, 3, and 4 in the "More information needed" column, and scores of 0 and 1 in the "Possible problem area" column. Another skill, also tested with 10 items but

having a mean difficulty of 70 percent, would have only scores of 7 through 10 in the "Satisfactory progress likely" column, with 3 through 6 in "More information needed," and 0, 1, and 2 in "Possible problem area."

On skills tested with small numbers of items, some adjustments in the above approach were made. These adjustments were necessary in order to avoid such absurdities as having zero scores in the "Satisfactory progress likely" column. Where only one item appeared for a skill, both scores of 0 and 1 were included in "More information needed."

Figure 2 provides an example of a portion of the Skill Summary Sheet. This sample can be used with Figure 1 to complete the process described in the following section.

Figure 2. Sample Section of the Skill Summary Sheet for Level 11, Form 8 of the Iowa Tests of Basic Skills. (Numerals represent possible raw scores.)

Test M-2: Mathematics Problem Solving			
	Possible Problem Area	More Information Needed	Satisfactory Progress Likely
1: single-step problems addition/subtraction	0 1 2	3 4 5	6 (7) 8
2: single-step problems multiplication/division	0	1 2 3	4 (5) 6 7
3: multiple-step problems combined use of basic operations	0 1 (2)	3 4 5	6 7 8 9 10 11 12

Description of the Interpretation Process

Using the two sample forms shown in Figures 1 and 2 together, a capsule form of the process can be seen. First the student would find the "Skill" row in Figure 1 and see that the first eight questions are coded for the major skill code "1" (one-step problems: addition and

subtraction). Next the student would count the pluses in the "Response" row for the items in this group (7 correct responses). Then, moving to Figure 2, the student would circle the 7 in the row of numbers beside skill area 1: single-step problems--addition and subtraction. This process would be repeated for skill areas 2 and 3 in the same manner, resulting in circles being drawn around the 5 in the middle row and around the 2 in the third row of Figure 2.

Since the student scored well into the "Satisfactory progress likely" column in skills 1 and 2 and somewhat lower in skill 3, it could be hypothesized that the student solves one-step problems easier than multiple-step problems. Since all four of the basic operations are involved in the first two skills, it could be further hypothesized that it is sorting out the elements of the multiple-step problems that is causing the difficulty in skill area 3 rather than computational errors alone.

Typical Time Allocations and an Outline for Conducting the Interpretations in Fourth, Fifth, and Sixth Grades

A typical pattern of the events and time allocations for conducting the interpretation in grades 4, 5, and 6 is given in Figure 3. These patterns were established in relatively small classes (less than 25 students) of about average overall ability. The time allocations need to be adjusted for accelerated or slower groups, but the overall change needed generally is not more than a few minutes.

The interpretation process followed the format given below in outline form:

I. Introduction.
   - A. Description of the test and reminder to the students of what the questions were like.
   - B. What the test measures and does not measure.
   - C. Reasons for taking tests.
   - D. Feelings that people have about taking tests.
   - E. Importance of finding out what the test results mean.

II. Reading the Pupil Item Response Record.
   - A. General organization of the sheet.
   - B. Through examples, teach students to read each row of information contained on the report.
   - C. Relate the data presented on the report back to what it represents in terms of test questions and skills tested.

Figure 3. Time Use in the Interpretation Process.

MINUTES	FOURTH GRADE	FIFTH GRADE	SIXTH GRADE
0	Hand out materials.	Hand out materials.	Hand out materials.
5	Introduction.	Introduction.	Introduction.
10	Explain Pupil Item Response Record.	Explain Pupil Item Response Record.	Explain Pupil Item Response Record.
15	Explain Skill Summary Sheets.	Explain Skill Summary Sheets.	Explain Skill Summary Sheets.
20	Guided practice.	Guided practice.	Guided practice.
25		Students complete Skill Summary Sheets with monitoring.	Students complete Skill Summary Sheets with monitoring.
30	Students complete Skill Summary Sheets with monitoring.		
35		Interpretation guidance for the completed Skill Summary Sheets.	Interpretation guidance for the completed Skill Summary Sheets.
40		Summary and general questions.	Summary and general questions.
45	Interpretation guidance for the completed Skill Summary Sheets.		
50	Summary and general questions.		

III. Using the Skill Summary Sheet to summarize performance.

A. Identify skill codes and model, through examples, how to match up skill codes on the Skill Summary Sheet to those on the report form.

B. Instruct students to count the number of correct answers (+'s) for each skill area tested and to circle the corresponding number on the Skill Summary Sheet.

IV. Students complete their Skill Summary Sheets.

A. Monitors should circulate through the group during this time, providing assistance for students who have problems and spot checking to be sure the students understand the process.

V. How to interpret and use the Skill Summary Sheets.

A. Define the three categories of performance.

1. Satisfactory progress likely: chances are good that the student has developed a working level of the skill tested.
2. Possible problem area: chances are good that lessons requiring this skill will be difficult for the student.
3. More information needed: performance was neither high enough nor low enough to make a well-founded guess about the development of the skill.

B. Along with the definition of the performance categories, encourage students to check with their teachers to be sure that the skill tested matches the curriculum sequence of the school. Some skills may be tested at earlier levels than they are taught in a particular curriculum. Low performance on those skills may be anticipated and, thus, should not be sources of undue concern for the student or teacher.

C. Look for skill areas within a test that deviate markedly from the general pattern of scores.

D. Caution students that high performance does not necessarily mean the skill has been mastered and that low performance does not mean that the student knows nothing about the skill.

VI. Summarize the process and encourage the students to discuss the results with their teachers and their parents.

PART III: THE IMPACT STUDY

As noted in the Introduction to this report, there is a substantial body of literature related to the need for providing feedback to students on the outcomes of tests they have taken. There are also a number of approaches to interpretation to be found in recent literature

(e.g. Bohning, 1979; Bradley, 1978; Cummings, 1981; Rudman, 1977). However, empirical evidence of the impact of these interpretation approaches on either teachers or students is lacking.

The major concern of this study was to determine whether the interpretation technique, reported by Cummings (1981) and described in Part II of this report, had any impact on student or teacher attitudes or knowledge about the standardized achievement test in use in their schools. The results of this study, in conjunction with the results of the two complementary studies presented in Parts IV and V of the report, constitute an initial attempt to validate this interpretation technique.

METHODS

Participants

Sampling was done by school building. Six building principals were contacted, and they agreed to participate in the impact study. The buildings were located in five public and one parochial district in eastern Iowa. These small city and rural districts ranged in total enrollments from 106 to 3,316 students. All fourth-, fifth-, and sixth-grade teachers and students in the buildings participated. Fall administration of the Iowa Tests of Basic Skills was part of their regular testing program. Because of the voluntary nature of the sampling it would be fair to assume that the building principals were positively inclined toward standardized testing and test interpretation. To the degree that principals' attitudes are related to teachers' additutes, the teachers would have been somewhat positive toward testing as well. The same reasoning applies to students.

This possible bias did not, however, affect the outcomes of the impact study because the classrooms within each building were randomly assigned to control and experimental groups. During the period from October 15, 1980, through January 25, 1981, all students received the test interpretation session with their teachers in attendance. The impact of the interpretation session was measured with a teacher questionnaire and a student questionnaire. The control teachers and students responded to their questionnaires just prior to the session, and the experimental teachers and students responded just after the

session. Table 1 presents the sample sizes. If one divides the number of students by the number of teachers in each cell of the table, disparate class sizes will be observed. There are two reasons for this. First, in one school district fifth and sixth graders were in combined classes; the students indicated their grade level on the questionnaire and were included in the impact study. However, the sixth-grade teachers were not included in the analysis. Second, several teachers invited guest teachers to attend the sessions and these teachers, in turn, responded to the teacher questionnaire and were included in the study.

Table 1. Participants in Study.

	Grade 4		Grade 5		Grade 6		Total	
	Teachers	Students	Teachers	Students	Teachers	Students	Teachers	Students
Control	5	112	10	292	9	252	24	656
Experimental	7	97	13	221	6	133	26	451
Total	12	209	23	513	15	385	50	1,107

Instruments

The teacher questionnaire. Twenty-one of the 35 items on the teacher questionnaire addressed teacher attitudes about the Iowa Tests of Basic Skills (ITBS), and 14 items measured knowledge of the ITBS (See Appendix A). The first 13 items asked teachers about their perceptions of the value of uses of ITBS results. On the basis of item content, five subscales were constructed with "extremely valuable" coded as 6, "very valuable" coded as 5, and so on to "not valuable" coded as 1. The items selected for each scale are listed below:

a) Reporting to others (mean of items 1, 2, and 3);

b) Use at the individual student level (mean of items 5, 7, 10, and 12);

c) Use at the student group level (mean of items 6, 8, 9, 11, and 13);

d) Instructional purposes (items 5, 6, 7, 10, 11, 12, and 13);

e) Administrative purposes (items 1, 2, 8, 9, and 11).

A utility scale was formed from items 14, 15, 16, 18, and 19 by coding the most negative response as 1 and the most positive response as 5 and computing the mean. This scale overlaps the content of several value scales, but the utility items were not included on the value scales because the response formats were different. Items 17, 20, and 21 were analyzed separately.

A 14-item knowledge scale was formed from items 22 to 35. Both multiple-choice and true-false questions were included, and the correct answers are indicated on the questionnaire in Appendix A. Most of the items asked about uses and interpretation of ITBS results; three items asked about interpretation of subskill results in general (items 25, 32, and 33). The median item difficulty (percent of correct answers) was .79, and the median corrected item to total score correlation was .25. Coefficient aplha was .54 which is respectable for a 14-item test which did not purport to measure a unitary trait.

In addition to the questionnaire, teachers were asked to complete a short evaluation form about the interpretation session. Both the experimental and control teachers completed the form after the session. The questions included on the form are described in the results section.

The student questionnaire. The student questionnaire (See Appendix B) also contained attitude and knowledge questions. Seven of the ten attitude items addressed attitudes about the ITBS and were analyzed in the impact study (three of the attitude items asked about teacher-made tests and are not discussed). The content of the attitude items did not lend itself to the formation of subscales, and they were analyzed separately. The most positive response was coded as 5 and the least positive response as 1.

The remaining 14 items assessed student knowledge about achievement testing. Nearly all of the items ask about purposes of the ITBS. All of the items are true-false, and the correct answers are listed in the questionnaire (Appendix B). For fourth-grade students, the knowledge questions were read aloud to the students, and students marked a machine-scorable answer sheet to record their responses. In the other

grades, and in all grades for the attitude questions, students read the questions and marked their responses on the answer sheet. The median item difficulty was .67 and median corrected item to total score correlation was .19. The coefficient alpha was .47.

As with teachers, both experimental and control students answered evaluation questions about the interpretation session just after the session. These questions are described under results.

Data Analysis

The main statistical tool used in data analysis was a 2 x 3 analysis of variance, with two levels of treatment group (experimental and control) and three levels of grade (fourth, fifth, and sixth grades). This procedure tested the effect of the experimental-control group membership (the effect of the skill interpretation session) and the effect of the grade level on the attitude and knowledge scale scores. Analysis of variance was also used to test effects upon individual attitude items.[1]

RESULTS

Impact of Interpretation on Teachers' Attitudes and Knowledge

Table 2 presents the results from the analysis of variance for teachers' attitudes and knowledge. None of the main effects of treatment or grade level were significant at the .05 level. One of the interactions was significant at the .05 level, but further analysis yielded no worthwhile interpretations. The interpretation of these data is straight forward--the short-term impact of the interpretation session on teacher attitude and knowledge was neligible.

In spite of the finding that short-term impact on teacher attitudes and knowledge was not found, the teachers' evaluations of the sessions indicated that the interpretation was perceived as worthwhile.

[1] Analysis of variance was performed on single items in spite of questions about the equal interval assumption for Likert-type items. This approach allowed all analyses to be reported in a consistent format. Where responses to individual items served as dependent variables, chi-square tests were also performed to test for relationships between treatment group and responses. In all cases, the chi-square results yielded interpretations which were equivalent to the results from analysis of variance.

Table 2. Results of Analysis of Variance for Teacher Attitudes and Knowledge.

	Grade Effect		Treatment Effect		Grade by Treatment Interaction	
	F	Prob	F	Prob	F	Prob
Perceived value of the ITBS results for:						
a) Reporting to others	1.56	.22	0.05	.82	0.26	.77
b) Use at the individual student level	0.66	.52	0.84	.37	2.09	.14
c) Use at the student group level	0.14	.87	0.46	.50	1.70	.20
d) Instructional purposes	0.27	.77	0.83	.37	3.17	.05
e) Administrative purposes	0.66	.52	0.07	.80	0.57	.57
Perceived utility of ITBS results	0.88	.42	0.93	.34	0.73	.49
Goodness of match between ITBS and curriculum (Item 17)	2.43	.10	0.07	.79	0.34	.72
Self-rated knowledge of ITBS (Item 20)	0.27	.77	0.96	.33	1.35	.27
Overall relative quality of ITBS (Item 21)	0.17	.85	1.12	.30	0.60	.55
Knowledge about the ITBS and interpretation of results (Items 22-35)	0.47	.63	0.07	.79	0.14	.87

Note: F = sequential F value as grade entered equation first, treatment entered second, and the interaction entered last.
Prob = probability of getting an F value equal to or larger than the observed F value under null conditions (where there is no effect of treatment or grade level).
Total sample sizes ranged from 47 to 50 because some of the items were omitted by some teachers.

In each of the sessions, the participating teachers responded to the five questions presented in Table 3. The percent of teachers selecting each response, for each question, is reported in the column to the left of the question.

Table 3. Evaluations of the Interpretation Sessions by Teachers.

Percent of Teachers Selecting Response	Evaluation Questions Asked and Response Options
	How difficult do you think the interpretation was for your students?
0	1. Too difficult
100	2. About right
0	3. Too easy
	How would you rate student interest in the interpretation session?
67	1. Very interested
29	2. Somewhat interested
4	3. Neutral
0	4. Somewhat bored
0	5. Very bored
0	6. Don't know
	Do you think the interpretation session will positively affect the students' test taking attitudes?
56	1. Yes
4	2. No
40	3. Not sure
	Do you think that the interpretation session and follow-up on it will result in improved teaching/learning?
65	1. Yes
2	2. No
33	3. Not sure
	Do you think that this type of interpretation session is worth continuing next year?
94	1. Yes
0	2. No
6	3. Not sure

One common criticism of test interpretation techniques is that the processes are complex and difficult for students to understand. The perceptions of the teachers who participated in the interpretation process under study here indicate that difficulty in understanding was not a problem at any of the three grade levels involved. In addition, the teachers felt that student interest in the session was very high. The questions about the effects of the sessions on students' test taking attitudes and regarding improved teaching/learning, addressed two long-term goals of the sessions. As might be expected, a sizable percentage of teachers were unsure about long-term effects. However, an even larger percentage (around 60 percent) felt that the session would positively affect students' test taking attitudes and result in improved teaching/learning. There was an interesting grade difference on the question about students' test taking attitudes; teachers of older students predicted more positive influence of the session on test taking attitudes than teachers of younger students. The percentages of yeses for the question were 33 percent for fourth grade, 50 percent for fifth grade, and 75 percent for sixth grade. Responses to the final question assessed overall evaluation, and it is apparent that the sessions were well received by the teachers.

Impact of Interpretation on Students' Attitudes and Knowledge

Table 4 presents the results from the analysis of variance for student attitudes and knowledge. Group means for those variables for which significant (at the .05 level) effects were obtained, are presented in Table 5.

The main effect of grade and interaction effects were found for both how well students like the Iowa Tests of Basic Skills, and for how difficult they perceive the tests to be. Fourth- and fifth-grade students liked the tests about equally well and were significantly more positive about the tests than sixth graders. In terms of the difficulty of the tests, fourth-grade students perceived the tests as most difficult, and fifth-grade students perceived them as easiest. The difference between fourth- and fifth-grade responses was significant. The other differences were not significant.

Table 4. Results of Analysis of Variance for Student Attitudes and Knowledge.

	Grade Effect		Treatment Effect		Grade by Treatment Interaction	
	F	Prob	F	Prob	F	Prob
Self-rated performance on ITBS (Item 1)	0.45	.64	4.84	.03	1.30	.27
Liking for ITBS (Item 3)	7.36	<.01	0.67	.41	8.63	<.01
Difficulty of ITBS (Item 4)	3.12	.05	0.60	.44	3.94	.0?
Anxiety toward ITBS (Item 7)	2.30	.10	0.15	.70	0.66	.52
Goodness of match between ITBS and curriculum (Item 8)	0.66	.52	0.62	.43	1.48	.23
Self-rated knowledge of ITBS (Item 9)	0.17	.85	0.01	.97	0.65	.52
Personal utility of ITBS results (Item 10)	0.94	.39	0.05	.83	1.67	.19
Knowledge of ITBS purposes (Items 11-24)	23.94	<.01	8.76	<.01	0.82	.44

Note: F = sequential F value as grade entered equation first, treatment entered second, and the interaction entered last.
Prob = probability of getting an F value equal to or larger than the observed F value under null conditions (where there is no effect of treatment or grade level).
Total sample sizes ranged from 1,104 to 1,107 because not all of the students completed all of the items.

Table 5. Group Means for Selected Student Attitude Items and Knowledge.

	Group	Grade 4	Grade 5	Grade 6	Total
Liking for ITBS (Item 3) I really like them = 5 I really hate them = 1	Experimental Control Total	2.89 3.49 3.20	3.21 3.09 3.14	2.89 2.93 2.92	3.04 3.20
Difficulty of ITBS (Item 4) Very hard = 5 Very easy = 1	Experimental Control Total	3.50 3.26 3.37	3.17 3.26 3.22	3.35 3.26 3.29	3.30 3.26
Self-rated performance on ITBS (Item 1) Quite high = 5 Quite low = 1	Experimental Control Total	3.50 3.25 3.37	3.37 3.26 3.31	3.31 3.29 3.30	3.38 3.27
Knowledge of ITBS purposes (Items 11-24) number of items correct	Experimental Control Total	9.39 8.73 9.04	9.65 9.14 9.36	10.24 10.00 10.08	9.77 9.40

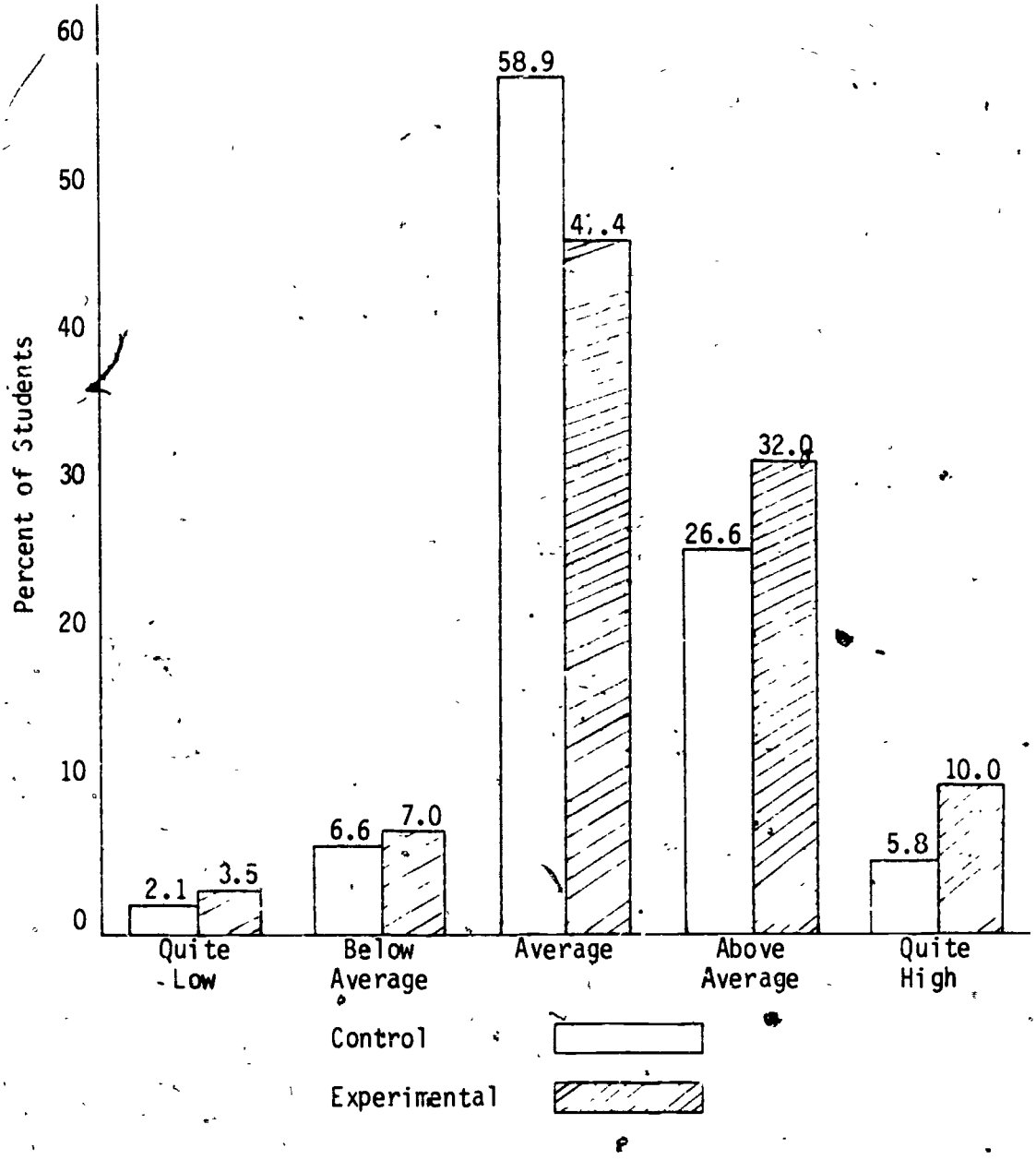
The interactions were such that, at the fourth grade, the control group liked the ITBS more and rated the difficulty lower than the experimental group. Across all grades, however, the control and experimental group were similar in their attitudes about the tests.

There were significant differences between the experimental and control groups on how well they thought they had done on the ITBS. At each grade level, the experimental group rated their performance higher than did the control. The group means are listed in Table 5, and Figure 4 illustrates the frequencies for all grades combined.

The experimental group also performed significantly higher on the 14-item knowledge test.[2] Table 5 shows that the experimental group was

[2] Although the students were randomly assigned to treatments, one could argue that the results of self-rated performance on the ITBS and the knowledge scale simply indicate that the experimental students were brighter than the control students. This would cast doubts on the significant treatment effects. In order to test this hypothesis, an analysis of covariance was performed. The grade and treatment effect upon the knowledge scale continued to be significant after adjusting for possible group differences in self-rated performance.

Figure 4. Frequency Distributions of the Experimental and Control Group for Item 3 (How well do you think you did on the Iowa Tests of Basic Skills this year?).



higher at each grade level. The main effect of grade was also significant with sixth graders scoring higher than fifth graders and fifth graders higher than fourth graders.

The students also answered questions specifically about the interpretation session--how interesting the session was, how confusing it was, and how helpful it might be in future learning. These evaluation questions were asked of all students just after the session and,

therefore, a breakdown for experimental and control groups was not appropriate. The total sample size ranged from 1,045 to 1,053, somewhat less than for previous results because two classes did not complete the evaluation questions. The results are presented in Table 6.

Table 6. Evaluations of the Interpretation Sessions by Students.

Percent of Students Selecting Response	Evaluation Questions Asked and Response Options
	Do you think the skill session was interesting?
54	1. Yes
29	2. No
17	3. Not sure
	Do you think the skill session was confusing?
20	1. Yes
55	2. No
25	3. Not sure
	Do you think that knowing your strong and weak areas will help you learn better?
77	1. Yes
7	2. No
16	3. Not sure

The students viewed the session as interesting, but their overall level of interest was not as high as their teachers perceived it. The grades did not significantly differ in reported interest level.[3] Although 100 percent of the teachers rated the difficulty of the session to be about right, only 55 percent of the students reported no problems with being confused by the session. There was a significant grade effect on this question. The sixth graders reported being less confused than the fourth and fifth graders. There was also a grade effect on the last question. Sixth graders were significantly higher than fifth graders, and fifth graders were significantly higher than fourth graders on their

[3]One way analysis of variance was conducted to investigate the effects of grade level on these items. Significant effects were further tested with Duncan's Multiple Range Test. Item responses were coded such that Yes = 3, Not sure = 2, and No = 1.

perceptions of the helpfulness of knowing their strong and weak areas. Across all grades the students were very positive about the effects on learning, more so, in fact, than the teachers.

CONCLUSIONS FOR THE IMPACT STUDY

The interpretation process had no immediate impact on either teacher attitudes/opinions or on teacher knowledge about the test, as assessed for the study. However, the evaluations of the interpretation sessions by teachers indicate that the sessions were positively received, were thought by most teachers to have the potential for positive effects in both future testing and teaching situations, and were considered by almost all teachers (94 percent) to be worth continuing next year. For students there was evidence of positive impact of the interpretation session. One important finding was the significant difference in knowledge found between students who had and those who had not been through the interpretation session. Cormany's (1974) study of attitudes toward standardized testing concluded that persons who felt they were well informed about the subject had more positive attitudes about it. If the increase in knowledge about the test, generated through the interpretation session, leads to feelings of being well informed (or better informed) about the test, then general attitudes toward the test may be improved over the long run.

The emphasis of this discussion of attitude change resulting from greater knowledge, however, must be on the long term potential effects, since no group differences were found for the short term effects of the interpretation session. The experimental and control groups did not differ in their ratings of several dimensions of the test, perceived utility of test results, or how nervous they felt before entering the testing situation.

The one opinion item which appeared to be directly affected by the interpretation session concerned how well the students thought they had done on the Iowa Tests of Basic Skills. Those students who participated in the interpretation process felt they had done better on the test than those who did not participate (see Figure 4). These results are particularly relevant in view of the frequent criticism that standardized tests may damage students' self-image. Figure 4

shows that student self-ratings, in general, clustered around the average rating with some skewness on the above average side. If the criticism were valid, the distributions would be skewed in the opposite direction. The findings further suggest that if test results are not interpreted with active student participation, students tend to rate themselves lower on the test and have a lower self-image of their abilities to achieve in school.

In summary, the interpretation process had an immediate impact on student knowledge about the Iowa Tests of Basic Skills and on the students' views about how well they had performed on the test. Other attitudes and opinions about the test were not immediately affected by the interpretation process. Further research on the long term effects on attitudes is needed.

PART IV: TEACHERS' PREDICTIONS OF STUDENT PERFORMANCE ON SUBSKILLS OF MATHEMATICS

A rarely asked, but critical, question for educators concerned about testing and the use of test results has been succinctly posed by Fitzgerald (1980): "Do tests provide much information about children's performance that teachers don't know by classroom observation?" (p. 216). If current practice is to be used as a guide in answering this question, then it might be said that teachers believe classroom observations are overwhelmingly the most useful of the ways of assessing students. Salmon-Cox (1981) recorded the finding that teachers most heavily depend on observation, perhaps, bringing into serious question Ebel's (1972, p. 49) assertion that "the majority of teachers and professors are keenly aware of the limited and unsatisfactory bases they ordinarily have for judging the relative achievement of various students and of the fallibility of their subjective judgments when based on the irregular, uncontrolled observations they can make in their classroom or office."

This study was an attempt to address one aspect of Fitzgerald's (1980) question. Teachers were asked to predict how their students would perform on the Mathematics subtests of the Iowa Tests of Basic Skills. The predictions, which were made both in terms of criterion-referenced (raw-score) performance and norm-referenced (percentile-rank)

performance, were later compared to the actual performance obtained on the tests by the students. The basic questions asked in this study were:

1. How highly correlated are the predicted and observed scores of students for subskill scores and for subtest total scores?
2. Do teachers tend to be accurate in their predictions of student obtained scores, and if over- or under-predictions occur, are they systematic?
3. Are there systematic differences in predictions that appear to be related to either grade, student sex, or overall mathematics achievement?
4. What relationships exist between predictions made on the basis of raw-score versus norm-referenced estimates?

METHODS

Participants

Forty-three fourth-, fifth-, and sixth-grade teachers, and a random sample of 374 of their students participated in this study. An average of between eight and nine students per teacher was selected, with a maximum set at ten students per teacher per class, in order to keep the number of predictions an individual would have to make within reasonable limits. The final student sample included 140 fourth graders, 117 fifth graders, and 117 sixtn graders. In each grade 55 percent of the students were males and 45 percent were females.

Of the teachers participating, 21 were fourth-grade, 14 were fifth-grade, and eight were sixth-grade teachers. Some of the sixth-grade teachers were mathematics teachers in a middle-school setting and, thus, made predictions for students from more than one class. The remaining teachers were responsible for the full range of teaching in self-contained classrooms. The participating teachers were drawn from six medium-sized to small school districts in eastern Iowa. Five public school districts and one private school participated.

Data Collection

Early in the fall semester of the 1980-81 school year, school districts and teachers were solicited for participation in the study. The students for whom predictions were to be made were selected randomly from class lists, and a Skill Summary Sheet for the appropriate

form and level of the test was prepared for each sampled student (see page 6 of this report for an example of the Skill Summary Sheet format). One to two weeks prior to the administration of the Iowa Tests of Basic Skills, the teacher was sent the Skill Summary Sheets for the sample of students from his/her class, along with directions for completing it. The directions instructed the teacher to "...circle the score you think the student named on the form will receive for each skill area listed." The three divisions of the Skill Summary Sheet were briefly explained, and an example was given with the directions. In addition, the teachers were asked to estimate the percentile rank that the student would obtain on each of the three mathematics subtests, Concepts, Problem Solving, and Computation. This prediction was done on a scale as shown in Figure 5. The request for a norm-referenced prediction on the basis of

Figure 5. Percentile Scale and Directions Used in Predicting Student Performance on Mathematics Subtests.

> Please estimate the percent of students in this grade statewide that this student is likely to score better than in Mathematics (subtest name) overall.
>
> Percent: 1 10 20 30 40 45 50 55 60 70 80 90 99
>
> (Circle the percent closest to your estimate.)

state, rather than national, norms was used because all of these school districts use state norms generated through their participation in the Iowa Testing Programs, and the teachers were accustomed to using state norms.

The predictions were made and returned to the project director before testing occurred in the regular school-wide testing program. The teachers had worked with the students a minimum of five weeks before the predictions were made.

Following the testing of the students, their Pupil Item Response Records were obtained from the schools and were used to calculate the raw score performance for each subskill tested for each student (see page 5 of this report for an example of the Pupil Item Response Record format). The actual percentile-rank performances of the students were

also obtained from their reports. These data, along with grade equivalents and the various predictions made by the teachers were coded and keypunched for the data analysis.

Data Analysis

The data analyses involved various comparisons between predicted and observed scores on the three subtests under study and the subskill categories comprising each subtest. The analyses included study of the effects of norm-referenced versus raw-score frameworks in making the estimates of performance, the accuracy of the predictions and over- or under-prediction, and the effects of student sex, grade level, and general mathematics achievement on the predictions.

The relationships between predicted and observed scores were established through correlational analysis. Raw-score estimates of performance and obtained raw scores were correlated, using Pearson product-moment correlations, for each subskill on the three mathematics subtests, and for total raw-score estimates for each subtest. The total raw-score estimates were calculated by summing the teachers' predictions on each subskill. In addition, correlations were computed for the predicted and observed percentile ranks for each subtest.

The accuracy of performance was analyzed in two ways. First, frequencies and percents were computed for discrepancy intervals, using predicted percent correct minus observed percent correct and ten unit intervals. Thus, with perfect prediction equal to zero, the interval indicating the greatest accuracy of prediction was the interval -4.9 to 4.9. The percent of students (i.e. predictions) for each subtest, falling within the discrepancy intervals was found from the frequency distributions. In addition to this descriptive approach, t-tests of differences between means were calculated for the subtest total scores and for the subskill scores. The t-tests for the raw-score estimates for subtest scores were based on the difference in percents between predicted and observed. For the norm-referenced calculations, the predicted and observed percentile ranks were converted to z scores, then handled in the same fashion as the raw-score estimates.

Correlations and t-tests were also computed for the subtest scores in analyses of the predictions based on student sex, grade level, and general matehmatics achievement level. To establish the mathematics

achievement level, the grade-equivalent scores of the students were summed across the three subtests (Concepts, Problem Solving, and Computation), and quartiles were computed for this grand-total mathematics score. The achievement groups were defined as: Low Achievers, the bottom 25 percent of students on this total score; Average Achievers, the middle 50 percent of students; and High Achievers, the upper 25 percent of students. These cutoff levels for low, average, and high were used because they are consistent with common practice in test interpretation for defining above average, average, and below average performance.

RESULTS AND DISCUSSION

Overall Correlations and Accuracy of Predictions

The correlations, based on estimates of raw-score performance for the subtests and the subskills are presented in Table 7. The correlations for the subtests are all in the 50s and are somewhat higher than the subskill correlations. This difference between subskill and subtest correlations was anticipated because of the decrease in reliabilities of both the test and the estimate when the shorter subskill scores are estimated. These correlations, in general, indicate a moderate amount of agreement in the rankings of the predictions and the actual scores obtained by the students.

To further assess the agreement between predictions and obtained scores, refer to Table 8. In Table 8, the percent of predictions about each of a student's subtest scores, falling within specified ranges of accuracy, is presented. Perfect prediction, which would be a discrepancy score of 0, is contained in the interval 4.9 to -4.9. The interval is essentially ten units wide, and represents the discrepancy between the predicted percent correct and the observed percent correct. Thus, it can be seen that 17.4 percent of the predicted Concepts scores were within $\pm$4.9 percent of being perfect predictions. It can further be seen that an additional 20 percent of the predictions were between 5 and 14.9 percent too high and that 17.9 percent of the students' Concepts scores were under-predicted by between 5 and 14.9 percent. It can be seen from the data in Table 8 that the tendency was for teachers to under-predict actual student performance in Concepts and to over-predict performance in Computation.

Table 7. Correlations Between Predicted and Observed Scores by Subtest and Subskill.

SUBTEST Subskill	Norm-Referenced Predicted VS Observed		Raw-Score Predicted VS Observed	
	n	r	n	r
CONCEPTS	319	.55	374	.53
Numeration			374	.38
Equations			374	.41
Whole Numbers			374	.40
Fractions			373	.32
Decimals/%			154	.32
Geometry & Measurement			373	.36
PROBLEMS	319	.57	374	.52
1-step, +, -			374	.34
1-step, x, ÷			374	.43
multiple-step			374	.45
COMPUTATION	319	.59	374	.58
Whole Number +			374	.39
Whole Number -			374	.45
Whole Number x			374	.54
Whole Number ÷			372	.51
Fractions +			234	.35
Fractions -			198	.26
Fractions x			117	.24
Decimals +			117	.42
Decimals -			117	.16*

*not significantly different from $r = 0$, $p \leq .01$.

Table 8. Percent of Students for Whom Raw-Score Predictions Were Accurate Within Specified Discrepancy Intervals.

Prediction Discrepancy Intervals* (predicted % correct - observed % correct)	Percent of Students Within Discrepancy Ranges		
	Concepts	Problems	Computation
≥35.0	3.7	5.6	5.1
25.0 to 34.9	4.0	7.5	7.2
15.0 to 24.9	9.6	11.2	14.2
5.0 to 14.9	20.0	12.8	21.7
4.9 to -4.9	17.4	25.4	24.1
-5.0 to -14.9	17.9	16.0	15.5
-15.0 to -24.9	13.6	11.8	6.7
-25.0 to -34.9	9.6	6.2	4.6
≤35.0	4.0	3.5	1.1

*perfect prediction = 0.

Table 9 carries the analysis of over- and under-prediction one step farther. In Table 9, the mean raw-score percents correct for predicted and observed are provided, along with t-tests of differences between the means. For the subtests, the conclusion reached in the consideration of Table 8 is supported, since a significant under-prediction was found for the Concepts subtest and a significant over-prediction was found for the Computation test.

It can also be seen in Table 9 that there is fairly consistent over-prediction for the subskills of the Computation subtest. This over-prediction might be attributable to a belief by teachers that the fundamental skills of mathematics computation have been achieved to a higher degree than they actually have. However, since the Mathematics Computation subtest of the Iowa Tests of Basic Skills is a more speeded test than any other part of the battery, it is likely that the speededness of the subtest contributed to the over-predictions. When the raw-score estimate is compared to the norm-referenced estimate, presented in Table 10, it is found that the relative standing (percentile ranks)

Table 9. T-tests of Differences Between Mean Predicted and Observed Raw-Scores by Subtest and Subskill.

SUBTEST Subskill	Mean Predicted % Correct	Mean Observed % Correct	$\bar{D}$	s_D	n	t	Significant* over (>) or under (<) prediction
CONCEPTS	54.2	56.4	.023	.195	374	-2.25	<
Numeration	61.5	60.7	.008	.246	373	.66	
Equations	60.8	57.1	.037	.300	374	2.39	>
Whole Numbers	58.7	58.5	.002	.266	374	.14	
Fractions	44.3	53.4	-.091	.302	373	-5.83	<
Decimals/%	44.3	39.2	.051	.399	154	1.58	
Geometry & Measurement	47.7	54.9	-.072	.265	373	-5.26	<
PROBLEM SOLVING	59.8	58.6	.012	.214	374	1.11	
1-step, +, -	74.9	67.7	.073	.254	374	5.53	>
1-step, x, ÷	55.7	52.5	.032	.301	374	2.03	
multiple-step	48.8	54.4	-.056	.248	374	-4.35	<
COMPUTATION	65.3	60.6	.047	.182	374	4.96	>
Whole Number +	82.5	78.4	.042	.224	374	3.60	>
Whole Number -	76.9	70.8	.061	.247	374	4.75	>
Whole Number x	62.6	61.6	.009	.236	374	.77	
Whole Number ÷	51.8	50.9	.008	.292	372	.54	
Fractions +	42.4	28.3	.140	.338	234	6.34	>
Fractions -	42.3	26.8	.159	.301	198	7.43	>
Fractions x	49.3	32.2	.171	.388	117	4.76	>
Decimals +	79.5	57.7	.218	.407	117	5.80	>
Decimals -	48.3	50.8	-.026	.440	117	-.63	

*$p < .05$.

of the students were more accurately predicted; thus, providing support to the hypothesis that speededness may have affected the raw-score estimates.

The subskill predictions in Table 9 for the Concepts and Problem Solving subtests, unlike those for the Computation subtest, are not consistent in their directionality. However, a pattern of over- and under-prediction can be distinguished. The subskills are presented in a roughly hierarchical order in Concepts and Problem Solving, and it can be seen that the pattern of over- or under-prediction followed an ascending order of skill complexity. That is, the lower level skills tended to be over-predicted and the more complex skills under-predicted.

Table 10 presents the t-test data for the norm-referenced predictions. $\bar{D}_z$ is the mean difference between predicted and observed percentile ranks that were converted to z scores. None of the t-tests were significant for the norm-referenced framework. This indicates

Table 10. T-tests of Differences Between Mean Predicted and Observed Percentile Ranks (converted to z scores) by Subtest.

Subtest	$\bar{D}_z$	s_D	n	t	p
Concepts	.038	.863	319	.78	.43
Problem Solving	-.096	.897	319	-1.92	.06
Computation	.031	.851	319	.66	.51

that the teachers showed less tendency to over- or under-estimate a student's relative performance (norm-referenced) than was true with the raw-score (criterion-referenced) predictions.

Correlations and Accuracy of Predictions by Sex, Grade, and Achievement Level

In an effort to determine which factors influenced the accuracy of the predictions about student performance, analyses were conducted on subgroupings based on student sex, grade level, and mathematics achievement as previously defined. The results of these analyses are presented here beginning with grade and sex, then turing to mathematics achievement.

Table 11 presents correlations for predicted and observed scores for both the raw-score and norm-referenced approaches by sex and grade. These correlations are fairly consistent across grades, sexes,

Table 11. Correlations Between Predicted and Observed Scores, by Sex and Grade.

		Raw-Score Prediction				Norm-Referenced Prediction			
	Sex	N	Concepts	Problems	Computation	N	Concepts	Problems	Computation
Grade 4	M	77	.58	.52	.59	63	.70	.72	.74
	F	63	.50	.48	.45	56	.59	.62	.55
	Total	140	.54	.50	.53	119	.65	.67	.65
Grade 5	M	64	.44	.41	.64	55	.41	.33	.50
	F	53	.48	.57	.58	49	.56	.64	.42
	Total	117	.46	.48	.62	104	.49	.48	.48
Grade 6	M	52	.58	.54	.67	43	.52	.63	.62
	F	65	.60	.62	.67	53	.54	.56	.66
	Total	117	.59	.59	.66	96	.52	.58	.65
Combined Grades	M	193	.54	.48	.62	161	.58	.55	.64
	F	181	.53	.55	.54	158	.52	.59	.53
	Total	374	.53	.52	.58	319	.55	.57	.59

subtests, and prediction approach (the two exceptions are the norm-referenced correlations for fifth grade, Problems boys versus girls, $z = -2.05$, $p < .05$; and for the same subtest when compared to the fourth-grade correlation, $z = 2.12$, $p < .05$*). This finding suggests

*These statistics were computed using the formula:

$$z = \frac{zr_1 - zr_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}$$

where: z = test of significance for differences in two correlation coefficients, zr = a z-transformation of the correlation coefficient, n = number of subjects in the sample.

that across most of the predictions of student performance the relationships between predicted and observed scores are similar.

In respect to the accuracy of predictions, however, fewer similarities across grades and sexes are to be found. Table 12 provides a summary of the patterns of over- and under-prediction of scores by sex and grade. Among the items of general interest in Table 12 are the consistent and significant over-predictions that occur for males in the Computation subtest under the raw-score approach. These large over-predictions for males and their small, but cumulatively significant, counterparts for females are what led to the previous finding of general, significant over-prediction for the Computation subtest. Also, of interest are the relatively accurate predictions for the Problem Solving subtest, and the inconsistencies that appear between the raw-score approach and the norm-referenced approach. While the number of significantly inaccurate predictions is approximately the same for the raw-score estimates (15) as for the norm-referenced estimates (16), in a number of cases the direction of the inaccuracy changes between the two methods--for example, grade 4 Computation or grade 5 Concepts. These shifts in over- and under-prediction may be an artifact of the different methods by which subtest scores were obtained (recall that in the raw-score approach the subtest total score was computed as a sum of the estimates for each subskill, whereas, the norm-referenced prediction was a single prediction for each subtest), or they may reflect some real difference in a teacher's ability to predict using the two different approaches. The current research design does not address this question.

Tables 13 and 14 present correlations and patterns of over- and under-prediction, respectively, for predictions for three mathematics achievement groupings, by grade. The correlations in Table 13 are all lower than those previously seen, and some are negative. This indicates that the modest precision that exists for the predictions when all achievement groups are ranked together is reduced, and in some cases lost altogether, when students are grouped along lines that are consistent with common test interpretation practices of identifying above average, average, and below average performance.

Table 12. Patterns of Over-Prediction and Under-Prediction by Sex and Grade, for Raw-Score and Norm-Referenced Predictions.

	n_{CR}	n_{NR}	Raw-Score						Norm-Referenced					
			Concepts		Problem Solving		Computation		Concepts		Problem Solving		Computation	
			Over	Under	Over	Under	Over	Under	Over	Under	Over	Under	Over	Under
Grade 4														
Male	77	63		X*		X	X*			X*		X*		X
Female	63	56		X	X		X			X*		X*		X*
Total	140	119		X*		X	X*			X*		X*		X*
Grade 5														
Male	64	55		X*	X		X*		X*		X		X*	
Female	63	49		X	X		X		X*		X		X	
Total	117	104		X*	X				X*		X		X*	
Grade 6														
Male	52	43		X	X		X*		X			X	X	
Female	65	53	X		X		X		X*			X		X
Total	117	96	X		X		X*		X*			X	X	
Combined Grades														
Male	193	161		X*	X		X*		X			X	X*	
Female	181	158		X	X		X*		X			X		X
Total	374	319		X*	X		X*		X			X	X	

*t-test of the mean difference between predicted and observed was significant $p < .05$.

Table 13. Correlations Between Predicted and Observed Scores, by Mathematics Ability Level and Grade.

	Ability Level					
	Low	(N)	Average	(N)	High	(N)
CRITERION-REFERENCED						
Concepts						
Grade 4	.13	(36)	.17	(68)	.31	(36)
Grade 5	.28	(31)	.04	(55)	.34	(31)
Grade 6	.16	(32)	-.02	(55)	-.16	(30)
Tot	.19	(99)	.12	(178)	.27**	(97)
Problem Solving						
Grade 4	.33*	(36)	.07	(68)	.27	(36)
Grade 5	.24	(31)	.03	(55)	.29	(31)
Grade 6	-.22	(32)	.12	(55)	.02	(30)
Total	.13	(99)	.07	(178)	.22*	(97)
Computation						
Grade 4	.12	(36)	.22	(68)	.12	(36)
Grade 5	.44**	(31)	.35**	(55)	.12	(31)
Grade 6	.00	(32)	.38**	(55)	.30	(30)
Total	.11	(99)	.31**	(178)	.17	(97)
NORM-REFERENCED						
Concepts						
Grade 4	.39*	(28)	.26*	(58)	.39*	(33)
Grade 5	-.42*	(28)	.11	(50)	.21	(26)
Grade 6	.16	(29)	-.23	(43)	-.15	(24)
Total	.29**	(85)	.04	(151)	.25*	(83)
Problem Solving						
Grade 4	.31	(28)	.40**	(58)	.26	(33)
Grade 5	.35	(28)	.05	(50)	.20	(25)
Grade 6	-.20	(29)	.17	(43)	.15	(24)
Total	.12	(85)	.18*	(151)	.23*	(83)
Computation						
Grade 4	.46**	(28)	.28*	(58)	.23	(33)
Grade 5	-.02	(28)	.29*	(50)	.14	(26)
Grade 6	.20	(29)	.20	(43)	.34	(24)
Total	.20	(85)	.26**	(151)	.24*	(83)

* r = o, p <.05.

** r = o, p <.01.

Table 14. Patterns of Over-Prediction and Under-Prediction by Mathematics Ability and Grade, for Raw-Score and Norm-Referenced Predictions.

Ability Level	n_{CR}	n_{NR}	Raw-Score						Norm-Referenced					
			Concepts		Problem Solving		Computation		Concepts		Problem Solving		Computation	
			Over	Under	Over	Under	Over	Under	Over	Under	Over	Under	Over	Under
Grade 4														
Low	36	28	X		X*		X*		X		X		X*	
Average	68	58		X		X	X*			X		X*		X
High	36	33		X*		X*		X		X*		X*		X*
Grade 5														
Low	31	28		X	X*		X*		X*		X*		X*	
Average	55	50		X	X		X*		X*		X		X*	
High	31	26		X*		X*	X			X		X*		X*
Grade 6														
Low	32	29		X	X			X	X*		X*		X	
Average	55	43	X*		X		X*		X*		X		X	
High	30	24		X		X	X*			X*		X*		X*

*t-test of mean difference between predicted and observed was significant $p < .05$.

Table 14 provides some fairly clear indications of the problems teachers have in estimating the performance of students at various achievement levels. Among the significantly inaccurate predictions shown in Table 14, 100 percent of those for low ability students, for both the raw-score and norm-referenced approaches, were over-predictions. Notably large percentages for the average ability students were also over-predictions (75 percent in the norm-referenced approach, and 100 percent in the raw-score approach). However, the direction of the inaccuracy shifted to under-prediction for the high ability students (in the norm-referenced approach, 100 percent of the significant predictions were under-predictions and in the raw-score approach 80 percent of the significant predictions were under-predictions). These findings suggest that for low and average achieving students, teachers tend to be overly optimistic about their students' subsequent performance on the test, and for high achieving students the teachers tend to be overly pessimistic.

SUMMARY OF THE STUDY OF TEACHERS' PREDICTIONS OF STUDENT PERFORMANCE ON SUBSKILLS OF MATHEMATICS

Fitzgerald (1980) asked whether tests provide information to teachers that they do not already know through classroom observations, and Salmon-Cox (1981) found that teachers' most frequently mentioned method of assessing their students was "observation." On the other hand, Ebel (1979) has contended that, "Most assessments of student achievement currently being made in our schools and colleges are ... highly subjective, uniquely individualistic, and unsystematic" (p. 11). Some testing programs in their interpretation materials emphasize that the "test data should confirm what a sensitive teacher already knows about students" (Prescott, et. al., p. 44), others, while recognizing that test results do not replace teacher judgment, appear to focus more on the discrepancies between teacher judgment and observed performance. The Iowa Tests of Basic Skills falls within the latter category.

Changes in education over time have led to increasingly formal settings, and teachers are now expected to deal with increasing numbers of students in their daily teaching activities (Chauncey & Dobin, 1963). This fact makes it increasingly difficult for teachers to meet Ahman

and Glock's (1975) challenge to know the student well enough to design appropriate educational programs to meet the objectives of instruction.

This study focused on the discrepancies between teachers' predictions of student achievement in mathematics and the subsequent actual achievement of the students involved. Teachers predicted both the raw-score performance and the relative standing (percentile-ranks) of randomly selected students in their classes. The predictions were based on the Iowa Tests of Basic Skills, which were in use in each of the classes.

In general, it was found that teachers were not very accurate predictors of student performance on the test. Further, it was found that some systematic biases appear to exist in the predictions. Males were frequently over-predicted, and the predictions for low and average ability students were also generally too optimistic. On the other hand, high ability students were generally under-predicted.

To the extent that subjective judgments of mathematics achievement are "better than" mathematics achievement as measured by the Iowa Tests of Basic Skills, the results of this study are diminished. However, the results of the study do indicate that in the absence of test results, many questions about individual student achievement in mathematics might never surface for a teacher. It is also indicated that biases which favor males, low ability, and average ability students could be brought into question through the appropriate interpretation of test results.

In respect to Fitzgerald's (1980) question concerning the additional information gained from tests over classroom observation, it appears that at least somewhat different information is often obtained from the two sources. Whether one source is better than the other is probably not a resolvable question. However, the fact that many discrepancies between teacher expectancies and test performance exist can lead to individuals, and perhaps certain subgroups of students, getting a closer look, as the teacher tries to explain the discrepancies for him- or herself. Ultimately, the questioning may lead to a more appropriate and more effective teaching program for the student, and the test will have served a valuable purpose.

PART V: RELATIONSHIPS BETWEEN THE RESULTS OF THE IOWA TESTS OF BASIC SKILLS, MATHEMATICS SUBTESTS, AND THE STANFORD DIAGNOSTIC MATHEMATICS TEST

One concern in validating a diagnostic interpretation technique for a test such as the Iowa Tests of Basic Skills (Hieronymus, et. al., 1978), is whether the survey test provides information that is similar to that obtained from recognized diagnostic tests. For this study, the Stanford Diagnostic Mathematics Test (Beatty, Madden, Gardner, & Karlsen, 1976) was used to compare student performances on the survey test and a diagnostic test.

It has been claimed that the results of the Iowa Tests of Basic Skills are "not useful for making decisions at the level of the individual child" (Harris, 1978, p. 57). On the other hand, the claim has been made for the Stanford Diagnostic Mathematics Test, that it "is an adequate test to identify the strengths and weaknesses of individual pupils in the areas covered" (Lappan, 1978, p. 437). This study was designed to provide both a structural and statistical assessment of the similarities between the Iowa Tests of Basic Skills subtests in Mathematics Concepts, Problem Solving, and Computation, and their counterparts of Concepts, Applications, and Computation on the Stanford Diagnostic Mathematics Test.

METHODS

Structural Comparisons of the Iowa Tests of Basic Skills and Stanford Diagnostic Mathematics Test

In preparation for the analyses of data collected under procedures described below, a structural analysis of the two test batteries was undertaken. This analysis involved a comparison between the tests to identify commonalities in the subskills tested, where the items assessing the subskills appeared in the respective test batteries, and to identify subskill categories that were tested on one of the batteries, but not on the other.

Item evaluators. Three item evaluators were assigned the task of reconciling the skill classifications on the two test batteries. These evaluators were: 1) a former test editor and testing consultant;

2) a graduate research assistant; and 3) the author of the mathematics tests of the Iowa Tests of Basic Skills. Each of these evaluators was familiar with the classification scheme used on the Iowa Tests, and each had previously reviewed the skills classification on the Stanford Diagnostic Mathematics Test.

The reconciliation process. Each of the three item evaluators independently reclassified each item on the Stanford Diagnostic Mathematics Test into its corresponding Iowa Tests of Basic Skills skill classification. The three independent classifications were then compared and discrepancies were reconciled, with the author making the final determination where disagreement existed. The result of this process was the reclassification and relabeling of the items of the Stanford Diagnostic Mathematics Test so that direct comparisons between the two test batteries could be made.

Comparisons of Performance on the Two Test Batteries Subjects

The students tested for this study were 288 fifth- and 260 sixth-grade students in a medium-sized (3,316 students, K-12) school district in eastern Iowa. The fifth-grade group was approximately 53.5 percent male and 46.5 percent female, and the sixth-grade group was approximately 46.5 percent male and 53.5 percent female.

Testing procedures. Both test batteries were administered during the fall semester of the 1980-81 school year. All students were administered the Iowa Tests of Basic Skills in the regular district-wide testing program in mid-September, 1980. Approximately two weeks later, participating students were administered the Stanford Diagnostic Mathematics Test. The tests were given in classroom groups, according to the directions specified in the manuals for the respective tests and under typical testing conditions. The tests were scored through the regular scoring services, provided by the Data Score Systems of Westinghouse Learning Corporation for both tests. Data tapes were obtained and merged for matched students, for whom complete item data were available on both tests.

Data analysis. For each subtest in its originally published form and for the reclassified subtests, defined through the structural analysis described above, classical test statistics were computed.

These included item p-values and discrimination indices, test means, standard deviations, and reliabilities (KR-20).

Using the test statistics generated and the student raw-scores, intercorrelations were computed between the published Concepts subtests, the Problem-Solving and Applications subtests, and Computation subtests. In addition, reliabilities of differences were computed, using the formula:

$$r_{dd} = \frac{r_{xx}S_x^2 + r_{yy}S_y^2 - 2r_{xy}S_xS_y}{S_x^2 + S_y^2 - 2r_{xy}S_xS_y}$$

Where:

r_{dd} is the reliability of the difference,

r_{xy} is the intercorrelation between the tests,

r_{xx}, r_{yy} are the respective reliabilities of the tests,

S_x^2, S_y^2 are the respective variances of the tests.

(Stanley, 1971, p. 385).

The same procedures for data analysis were repeated for the restructured tests, which included the three subtests (in restructured form) from the first analysis, plus a new subtest, called Graphs and Tables, which was defined in the structural analysis of the tests.

RESULTS AND DISCUSSION

Structural Analysis and Item Reclassification

Both the Stanford Diagnostic Mathematics Test and the Iowa Tests of Basic Skills have subtests in Mathematics Concepts, Applications or Problem-Solving, and Computation. However, major differences were observed in the item types and contents found under the various subtest labels.

It was found that nine items appearing in each Applications subtest of the Stanford Test were items involving the reading of graphs and tables. These items were similar to a subset of items appearing in the Work-Study Skills, Visual Materials subtest of the Iowa Tests. Therefore, for both batteries, a new subtest was defined and called, Graphs and Tables. These new subtests were included in the data analysis for the reclassified tests.

Aside from the inclusion of items from the Work-Study Skills portion of the Iowa Tests, the three Iowa Mathematics subtests were held intact for all of the analyses. The reclassification of the Stanford Test items is summarized in Table 15. The table shows, for example, that of the 36 grade 5 items in the Concepts subtest, all were reclassified into Concepts skill categories by the item evaluators. However, an additional ten items, originally published as Applications items, were reclassified into Concepts items, and nine items originally appearing in the Stanford Computation subtest were reclassified as Concepts items. The result of casting the Grade 5 Stanford items into Iowa Tests of Basic Skills classifications was a newly defined Concepts subtest of 55 items. This compares to a published version of the Stanford Concepts subtest of 36 items.

The general pattern of the reclassifying of the Stanford items yielded a considerably heavier emphasis on mathematics Concepts than might be assumed by looking at the subtest titles. Another striking outcome of the reclassification was the drop in emphasis shown for measuring skills in solving story problems. Even if the story problems and Graphs and Tables were combined, the resulting Applications subtest would be shorter by more than a third, because of the Concepts items imbedded in it. Limitations in the Stanford Diagnostic Mathematics Test item classification schemes have been previously noted (Sowder, 1978), but these limitations have not been shown so dramatically as they appear here.

Only three items, of the 231 making up the two levels of the Stanford Diagnostic Mathematics Test, were of a type that had no counterpart on the Iowa Tests of Basic Skills. Therefore, although there appears to be some difference in the emphasis of the various mathematics skills tested between the two tests (as measured by the number of items allocated to different skill categories), the two test batteries do measure similar skills. In general, the Stanford samples a somewhat narrower content domain, but includes a greater number of items for each of those skills represented. These are the kinds of differences one would expect between a "diagnostic" and a "survey" battery with approximately the same total number of items in each battery. The differences in emphasis in specific subskills are illustrated in Table 16.

Table 15. Number of Stanford Diagnostic Mathematics Test Items Reclassified to Comparable Iowa Tests of Basic Skills Classifications.

	Concepts		Applications		Computation		Total # of Items in the Reclassified Tests	
	Grade 5	Grade 6	Grade 5	Grade 6	Grade 5	Grade 6	Grade 5	Grade 6
# of items in the published tests	36	36	30	33*	48	48	--	--
# of items in the reclassified tests								
Concepts	36	36	10	12	9	6	55	54
Applications	0	0	11	9	0	0	11	9
Computation	0	0	0	0	39	42	39	42
Graphs and Tables	0	0	9	9	0	0	9	9

*No equivalent items or item classifications appeared in the Iowa Tests of Basic Skills for three of the items in the Stanford Applications subtest.

Table 16. Comparison of Items Allocated to Specific Subskills on the Iowa Tests of Basic Skills and the Stanford Diagnostic Mathematics Test for Grades 5 and 6.

Subtest: Mathematics Concepts	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Numeration, number systems, and sets	9	14	9	20
Counting and number series	·	3	·	4
Place value and expanded notation	3	6	3	5
Properties of number systems	3	4	4	11
Subsets of number systems	1	1	1	·
Sets	2	·	1	·
Equations, inequalities, and number sentences	4	6	5	9
Operational and relational symbols	1	·	2	·
Solution of number sentences	3	6	3	9
Whole numbers; Integers	6	13	7	13
Reading and writing	1	3	·	3
Relative values	·	2	·	2
Terms	2	2	2	·
Fundamental operations: Number facts	1	·	2	2
Fundamental operations: Ways to perform	1	4	1	3
Fundamental operations: Estimating results and rounding	1	2	2	3
Fractions	8	6	7	3
Part of a whole and partitioning of a set	1	2	2	2
Relative values	1	2	1	1
Equivalent fractions	3	1	1	·
Terms	1	1	1	·
Fundamental operations: Ways to perform	1	·	1	
Fundamental operations: Estimating results	·	·	1	·
Ratio and proportion	1	·	·	·

Table 16. Continued.

Subtest: Mathematics Concepts	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Decimals, currency, and percent	5	3	·	2
Reading and writing	1	·	·	·
Relative values	1	2	·	2
Fundamental operations: Estimating results and rounding	1	·	·	·
Equivalence: Decimals, fractions, and percents	1	1	·	·
Probability and statistics	1	·	·	·
Geometry and measurement	8	·	9	7
Measurement: Quantity, time, and temperature	1	·	2	1
Measurement: Length and weight	3	·	3	3
Recognizing types and parts of geometric figures	1	·	2	3
Area and perimeter of plane figures	2	·	1	·
Use of geometric figures in description and proof	1	·	1	·

Subtest: Mathematics Problem Solving	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Single-step problems: Addition - Subtraction	9	1	11	3
Currency	1	1	3	1
Whole numbers	7	·	6	2
Fractions, decimals, percents	1	·	2	·
Single-step problems: Multiplication - Division	7	2	6	4
Currency	2	·	·	·
Whole numbers	4	2	5	4
Fractions, decimals, percents	1	·	1	·

Table 16. Continued.

Subtest: Mathematics Problem Solving	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Multiple-step problems: Combined use of basic operations	13	6	10	4
Currency	8	3	6	4
Whole numbers	5	3	4	·
Fractions, decimals, percents	·	·	·	·

Subtest: Mathematics Computation	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Whole number	28	33	38	39
Addition	5	3	10	6
Subtraction	5	6	9	12
Multiplication	9	12	12	12
Division	9	12	7	9
Fractions	13	3	7	·
Addition	4	·	3	·
Subtraction	6	2	4	·
Multiplication	3	1	·	·
Division	·	·	·	·
Decimals	4	6	·	·
Addition	2	2	·	·
Subtraction	2	1	·	·
Multiplication	·	3	·	·
Division	·	·	·	·

Subtest: Graphs and Tables	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Reading amounts	1	3	2	3
Using the scales on bar and line graphs	1	3	2	3

Table 16. Continued.

Subtest: Graphs and Tables	ITBS Grade 6	SDMT Grade 6	ITBS Grade 5	SDMT Grade 5
Comparing quantities	4	6	8	6
Determining rank		2	3	3
Determining differences between amounts	4	4	4	3
Determining ratios			1	

Another way of determining whether the two batteries measure mathematics skills in comparable ways is through statistical analysis of the performance of students. Such an analysis is discussed in the following section.

Comparisons of Performance on the Iowa Tests of Basic Skills and the Stanford Diagnostic Mathematics Test

The analysis of performance involved two stages, one for the tests as they were published, and a parallel investigation for the tests as they were reclassified through the structural analysis. In each case, item p-values were determined and discrimination indices were computed. Table 17 presents the mean p-values and mean item-total correlations for the various subtests. It can be seen from the table that the Stanford Diagnostic Mathematics Test is a considerably easier test than the Iowa Tests of Basic Skills. This finding is not surprising, since the Stanford Test has been specifically designed to discriminate well among the lower achieving students on the skills tested. The Stanford also shows somewhat higher mean biserial discrimination indices than the Iowa Tests. This difference in discrimination values is most likely caused by the greater content homogeneity of the Stanford tests.

The reliabilities of the various subtests are presented in Table 18 on page 48, along with the intercorrelations between similar subtests on the Stanford and the Iowa batteries. In addition, estimates of the reliability of differences between subtest scores are given.

Table 17. Mean p-values and Biserial Correlations for the Published and Reclassified Tests.

Subtest	ITBS		SDMT (Published)		SDMT (Reclassified)	
	$\bar{p}$	mean r_{bis}	$\bar{p}$	mean r_{bis}	$\bar{p}$	mean r_{bis}
Grade 5						
Concepts	59.27	.487	78.64	.589	77.49	.616
Problems/Applications	62.44	.574	73.53	.623	67.64	.625
Computation	60.64	.548	82.58	.634	83.51	.600
Graphs and Tables*	57.60	.417			82.00	.679
Grade 6						
Concepts	57.12	.559	55.58	.599	59.18	.600
Problems/Applications	59.86	.602	70.58	.635	71.33	.676
Computation	60.91	.612	73.38	.654	72.88	.646
Graphs and Tables*	41.20	.420			80.89	.713

*The Graphs and Tables subtest, for both ITBS and SDMT, is a created subtest using subtsets of items from the Work-Study Skills, Visual Materials Test of ITBS and from the Applications Test of the SDMT.

Table 18. Number of Items and KR-20s by Test; and Intercorrelations and Reliabilities of Difference for Like Tests for Each Grade Level and for the Published and the Reclassified Forms of the Tests.

	Grade	N	Concepts		Problems		Computation		Graphs And Tables	
			ITBS	SDMT	ITBS	SDMT	ITBS	SDMT	ITBS	SDMT
PUBLISHED TESTS										
No. of Items (k)	5	288	37	36	27	30	45	48	---	---
KR-20			.82	.85	.84	.86	.88	.89	---	---
Intercorrelation			.674		.740		.629		----	
Reliability of Difference			.525		.403		.701		----	
No. of Items (k)	6	260	40	36	29	33	45	48	---	---
KR-20			.88	.88	.87	.88	.90	.92		---
Intercorrelation			.844		.771		.770		----	
Reliability of Difference			.251		.449		.614		----	
RECLASSIFIED TESTS										
No. of Items (k)	5	288	37	55	27	11	45	3[illegible]	10	9
KR-20			.82	.90	.84	.72	.88	.8[illegible]	.44	.70
Intercorrelation			.747		.720		.[illegible]17		.363	
Reliability of Difference			.434		.213		.663		.321	
No. of Items (k)	6	260	40	54	29	9	45	42	5	9
KR-20			.88	.92	.87	.74	.90	.91	.39	.74
Intercorrelation			.858		.683		.756		.345	
Reliability of Difference			.299		.394		.610		.333	

The subtest reliabilities are generally respectable for both test batteries, ranging between .70 and .92, except for the Graphs and Tables subtests created for the Iowa Tests of Basic Skills. Further, for the longer subtests overall, the reliabilities range above .80.

With the small numbers of items (ranging from 5 to 10) in the Graphs and Tables subtests, relatively low reliabilities were found. Since the low reliabilities associated with the subtests restricted the intercorrelations between the tests and the restricted intercorrelations in turn inflated the reliabilities of differences for these tests, the results are presented, but not discussed.

The interpretation of the reliabilities of differences should help in determining whether the two test batteries are functioning differently in a statistical sense. In one respect, the intercorrelations presented in Table 18 could be viewed as concurrent validity coefficients. That is, as estimations of the degree to which the two test batteries measure the same attributes. From that perspective, the correlations are reasonably high. However, this raw correspondence and its interpretation can be enhanced by study of the reliabilities of difference. In this case, the reliabilities of differences represent a measure of the stability of the difference scores observed between, for example, the two Concepts subtests. The higher the reliability of the difference, the more stable that difference is assumed to be. In other words, "real" differences, rather than differences attributable to error, are associated with high reliabilities of differences. When high intercorrelations and low reliabilities of differences exist, the subtests can be said to not be measuring statistically unique attributes.

The interpretation of reliabilities of differences can be approached in the same way a test user would approach interpreting the reliability of a test (Schreiner, Hieronymus, & Forsyth, 1969). However, it should be clear that high reliabilities of difference can be obtained only when two highly reliable measures, with low intercorrelations have been used (Stanley, 1971).

The reliabilities of differences presented in Table 18 do not provide a definitive answer to the question wnether the Iowa Tests and the Stanford Test measure the same attributes. In fact, some of the results are indeed surprising.

First among the surprises contained in Table 18 is the fact that for the published Concepts subtests, the reliability of differences was markedly different for the fifth- and sixth-grade groups. This finding indicates that the similarities in the Concepts subtests are greater at the sixth-grade level than at the fifth. This difference was somewhat reduced under the analysis of the reclassified tests and, perhaps, supports the earlier contention that reclassification of Stanford items led to structuring reasonably comparable tests, in terms of skill coverage. The same phenomenon, of reduced difference between fifth- and sixth-grade results appeared in the Computation subtest, and may further support the preceding statement. In general, the reliabilities of differences among the published versions of the Problems and Computation subtests were comparable between the two grades.

The second surprise was the high reliabilities of differences for the Computation subtests, relative to either the Concepts or Problems subtests. This finding suggests that the greatest likelihood of the Iowa and Stanford tests measuring different attributes is found in the Computation subtests. While this finding is counter-intuitive, it may be explained in part through a consideration of the speededness of the two tests. Since the Iowa Computation subtest is relatively speeded, but the Stanford Computation subtest is essentially a power test, the difference may be, in part, explained. The difference may not be one of computation skills measured, but one of speed and accuracy, versus accuracy alone.

Third among the surprises was that the differences between the published forms and the reclassified forms of the tests were, in general, relatively small. The restructuring of the Stanford Tests according to the Iowa's skills classification scheme did have substantial impact on the Problems (Applications) subtest. However, at least a portion of that impact can be attributed to the small number of "story" problems left after the reclassification, and the consequent lowering of the reliabilities of the subtests.

It should be noted that all of the reliabilities of differences are inflated to an unknown degree. The intercorrelations between the Stanford tests and Iowa tests include day-to-day sources of variation

in pupil performance, while the Kuder-Richardson 20 reliability coefficients for these tests do not. The use of more appropriate parallel forms reliabilities in estimating the reliabilities of differences would have lowered the obtained values substantially. The effect of taking into consideration different sources of error in computing such reliabilities is documented in the Manual for Administrators, Supervisors, and Counselors of the Iowa Tests of Basic Skills (Hieronymus & Lindquist, 1974, pp. 71-73).

SUMMARY OF THE RELATIONSHIPS BETWEEN THE IOWA TESTS OF BASIC SKILLS AND THE STANFORD DIAGNOSTIC MATHEMATICS TEST

The study of the relationships between the Iowa Tests of Basic Skills and the Stanford Diagnostic Mathematics Test was carried out on two levels. First, a structural match between the two test batteries was undertaken, and second, a correlational study of the intercorrelations of "like" subtests and the reliabilities of differences, involving 288 fifth- and 260 sixth-grade students, was conducted.

The structural analysis led to conclusions that a number of items appearing in the Applications subtest of the Stanford were, in fact, according to the Iowa skills classification, measuring mathematics concepts. Additionally, there were graphs and tables items in the Applications test that corresponded to items from the Work-Study Skills area of the Iowa Tests, and computation items that were considered to be measuring concepts as defined for the Iowa Tests. In general, however, almost all (99 percent) of the items on the Stanford Diagnostic Mathematics Test were found to have equivalent counterparts on the Iowa Tests of Basic Skills. The conclusion drawn was that the two test batteries measure essentially the same mathematics skills.

The study of intercorrelations and reliabilities of differences, however, did not lead to as clear cut a conclusion. The intercorrelations, while creditably high for purposes of looking at construct validity, led to reliabilities of differences that were also higher than would be expected if the tests were measuring the same attributes in the same way. Although, as noted, these reliabilities were somewhat inflated since the KR-20 reliabilities and intercorrelations used in their computation contained different sources of error variance.

The level that a reliability of difference must attain to be significant is an interpretation problem, however, not a statistical problem. For purposes of work with individual student scores, the reliabilities of differences could, therefore, be considered to be generally low enough to be attributable to measurement errors. Thus, the conclusion that the same skills are being measured was tentatively supported.

PART VI: SUMMARY OF THE PROJECT AND CONCLUSIONS

This project was implemented to study the need for, feasibility of, and impact of an interpretation technique designed for use with a standardized achievement test. The interpretation technique was developed for use with the Iowa Tests of Basic Skills, and the studies reported were specific to that test. However, the similarities between the reporting systems of the Iowa Tests and other major standardized, achievement test batteries, and the general principles applied in the development of the interpretation technique make the approach generalizable to other tests for which the subskills measured are fairly well defined.

The main focus of the project was to assess the impact of the interpretation on students and teachers. This study was reported in Part III: The Impact Study. However, two other important questions were addressed through the project. The first of these, addressed in Part IV: Teachers' Predictions of Student Performance on Subskills of Mathematics, dealt with the accuracy of teachers' expectations of student performance on the tests. The importance of this study was its focus on the commonly held belief that the subjective observations that teachers make in their day-to-day classroom activities lead to the formation of accurate assessments of student skill development.

The second important question studied, beyond the impact of interpretation, was whether the Iowa Tests of Basic Skills, Mathematics subtests, were comparable in design and function to a widely recognized "diagnostic" mathematics test. This study, reported in Part V: Relationships Between the Results of the Iowa Tests of Basic Skills, Mathematics Subtests, and the Stanford Diagnostic Mathematics Test,

addressed the feasibility of the diagnostic interpretation of the "survey" test results. This study was important in establishing or refuting the basic premise upon which the interpretation technique was developed.

The findings of the three studies incorporated into this project lead to a conclusion that there is a need for the interpretation of the results of tests administered in school-wide testing programs. At least two bases for this conclusion were found. First, students who have been through an interpretation process feel that they have done better on the test than students who have not had the test results interpreted to them. Presumably, it can be inferred from this finding that students will then feel "better" about themselves and their skill development. Secondly, the act of interpretation should raise important questions for the teachers as discrepancies between expectations and actual performance occur. This should benefit both students and teachers, as reasons for the discrepancies between the students' behaviors and the teachers' expectations are explained. The benefit for teachers should be an opportunity to: 1) reassess their expectations for certain students; and 2) examine some of their biases about the performance of certain subgroups in the subject areas tested. The benefit for students should be a better educational process borne out of better expectations for themselves and more appropriate expectations from their teachers, regardless of the student's sex or overall achievement level.

There was modest support for providing "diagnostic" interpretation of the "survey" test. This support came through the study of relationships between the Iowa Tests and the Stanford Diagnostic Mathematics Test. One weakness of this study may be found in the definition of a "diagnostic" test, and in this case, the Stanford was used primarily because it is promoted as a diagnostic instrument. This, previously challenged use for survey tests among testing professionals, but often practiced use among teachers and counselors, still is the source for the most serious cautions in the interpretation process presented.

The problem that arises in the "diagnostic" interpretation of tests like the Iowa Tests of Basic Skills is that highly related subskills become the focus of attention in the interpretation. The

subskills of mathematics, for example, yield high intercorrelations in part because the student that achieves well in one area of mathematics is likely to also achieve well in other mathematics skills. These high intercorrelations lead to low statistical reliabilities of difference between the subskills. This means that profiles of scores, observed at one testing period, may not be stable if the test is administered again. In these instances, it can be argued that a measure on one skill is indicative of the student's ability on the other, and any difference in the student's profile is attributable to measurement error. This statistical argument ignores the qualitative differences between the sets of items, but it is, none the less, an important consideration in the use of an interpretation technique such as the one studied here.

This area of profile analysis on achievement test results is one that deserves a great deal more attention than it has received. Most of the existing research in the area has been done in reading comprehension, under some fairly restrictive assumptions about what readers are like. The area could benefit from studies that replicate interpretation practices that more closely resemble those that occur in practice and through extension of the investigations to other subject areas represented on tests.

Another important caution regarding this interpretation technique is that it sets the test results into fairly concrete, easy to understand terms (i.e. the raw scores for the subskills tested). While this approach demystifies the test interpretation process to some extent, it also could lead to overconfidence or overinterpretation of the scores. It is important for the user to keep these performances in perspective just as they should any other test score. Although this caution is an important one, the results of the impact study suggest that this may be an unfounded concern about the process. Very few short term changes in attitudes about the test or its uses were shown to be related to the interpretation process.

In summary, three studies were conducted in relation to an interpretation process designed to actively involve students in the interpretation of their performance on a standardized achievement test. The studies provided support for the need, the feasibility, and a few

important outcomes of the interpretation technique. One of the important, but unstudied, underlying aspects of this interpretation technique is that the student becomes an active, rather than a passive, recipient of test results. The impact of these two different approaches to providing test results is another area for future study.

REFERENCES

Ahman, J. S. & Glock, M. D. Evaluating Pupil Growth: Principles of Tests and Measurements. Boston: Allyn and Bacon, Inc., 1975.

American Personnel and Guidance Association. Responsibilities of Users of Standardized Tests. Falls Church, Virginia: APGA Publication Sales, 1980.

American Personnel and Guidance Association & National Council on Measurement in Education. The responsible use of tests: A position paper of AMEG, APGA, and NCME. Measurement and Evaluation in Guidance, 1972, 5, 385-388.

Beatty, L. S., Madden, R., Gardner, E. F., & Karlsen, B. Stanford Diagnostic Mathematics Test. New York: Harcourt Brace Jovanovich, Inc., 1976.

Bohning, G. A profile for communicating achievement test results to children. Elementary School Guidance and Counseling, 1979, 13, 256-260.

Bradley, R. W. Person-referenced test interpretation: A learning process. Measurement and Evaluation in Guidance, 1978, 10, 201-210.

Bradley, R. W. Meeting test consumers' needs through interpretation. A paper presented at the annual meeting of the American Personnel and Guidance Association, St. Louis, Missouri, April, 1981.

Buros, O. K. Fifty years in testing: Some reminiscences, criticism, and suggestions. Educational Researcher, 1977, 6, 9-15.

Chauncey, H. & Dobbin, J. E. Testing: Its Place in Education Today. New York: Harper & Row, Publishers, 1963.

Cormany, R. B. Faculty attitudes toward standardized testing. Measurement and Evaluation in Guidance, 1974, 7, 188-194.

CTB/McGraw Hill. California Achievement Tests: Test Coordinator's Handbook, Preliminary Edition. Monterey, California: CTB/McGraw-Hill, 1977.

Cummings, O. W. Student-centered test interpretation: An active technique. The School Counselor, 1981, 28, 267-272.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

Ebel, R. L. Using tests to improve learning. Arithmetic Teacher, 1979, November, 10-12.

Fitzgerald, S. M. What are the effects of tests? Childhood Education, 1980, February/March, 216-217.

Harris, L. A. Iowa Tests of Basic Skills, Forms 5 and 6, Review. In O. K. Buros (Ed.). The Eighth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1978, 55-57.

Hieronymus, A. N. & Lindquist, E. F. Iowa Tests of Basic Skills: Manual for Administrators, Supervisors, and Counselors. Boston: Houghton Mifflin Company, 1974.

Hieronymus, A. N., Lindquist, E. F., & Hoover, H. D. The Iowa Tests of Basic Skills. Boston: Houghton Mifflin Company, 1978.

Hieronymus, A. N., Lindquist, E. F., & Hoover, H. D. Iowa Tests of Basic Skills: Teacher's Guide for Administration, Interpretation, and Use (Levels 9-14). Boston: Houghton Mifflin Company, 1979.

Kirkland, M. C. The effects of tests on students and schools. Review of Educational Research, 1971, 41, 303-350.

Lappan, G. Stanford Diagnostic Mathematics Test Review. In O. K. Buros (Ed.). The Eighth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1978, 436-437.

Lyman, H. B. Know the score before the game begins. Measurement and Evaluation in Guidance, 1974, 7, 150-156.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. Metropolitan Achievement Tests: Teacher's Manual for Administering and Interpreting. New York, New York: The Psychological Corporation, 1978.

Rudman, H. C. The standardized test flap. Phi Delta Kappan, 1977, November, 179-185.

Salmon-Cox, L. Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, 1981, 62, 631-634.

Schreiner, R. L., Hieronymus, A. N., & Forsyth, R. Differential measurement of reading abilities at the elementary school level. Reading Research Quarterly, 1969, Fall, 84-99.

Sowder, L. Stanford Diagnostic Mathematics Test Review. In O. K. Buros (Ed.). The Eighth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1978, 437-438.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.). Educational Measurement (2nd Edition). Washington, D. C: American Council on Education, 1971, 356-442.

APPENDIX A: TEACHER SURVEY

GENERAL DIRECTIONS: This survey consists of several types of questions to assess your opinions and knowledge of the Iowa Tests of Basic Skills. Please mark your answers directly on this survey. Specific directions are given with each type of question.

DIRECTIONS: Below is a list of possible uses for Iowa Tests of Basic Skills results. Using the following code, please give your opinion on the value of each use by checking (✓ or x) the appropriate column.

1= Extremely valuable for this use.
2= Very valuable for this use, the same use could be met in other ways only with great difficulty.
3= Valuable, the same use could be met in other ways with some effort.
4= Somewhat valuable, the same use could be met in other ways without much difficulty.
5= Minimally valuable, the test results are useful for "added information" but not for meeting the objective.
6= Not valuable, the test results create problems or detract from other information that could be better used to meet the objective.

		Extremely Valuable	Very Valuable	Valuable	Somewhat Valuable	Minimally Valuable	Not Valuable
1.	Reporting to local news media						
2.	Reporting to boards of education						
3.	Reporting to parents						
4.	Screening of special education students						
5.	Planning instruction for individual students						
6.	Planning instruction for groups of students						
7.	Comparing individual scores with performance of a state or national peer group						
8.	Evaluating specific teaching procedures or methods						
9.	Comparing classes within a school						
10.	Measuring individual growth from year to year						
11.	Identifying system-wide strengths and weaknesses						
12.	Identifying individual strengths and weaknesses						
13.	Grouping students for specific instruction						

DIRECTIONS: Please answer the following opinion questions by putting the number of your answer in the blank at the right of each question.

14. How relevant are the results of the ITBS to your work with Students? (1) Not at all relevant (2) not very relevant (3) Somewhat relevant (4) Very relevant (5) Extremely relevant

15. How useful are the results of the ITBS in identifying strong or weak points in the curriculum? (1) Not at all useful (2) Minimally useful (3) Useful to some extent (4) Useful to a great extent (5) Useful to a very great extent

16. How useful are the results of the ITBS in discussing future instructional plans with individual students? (1) Not at all useful (2) Minimally useful (3) Useful to some extent (4) Useful to a great extent (5) Useful to a very great extent

17. How closely do the skills tested on the ITBS match the skills in the curriculum you actually teach? (1) Very high match (2) High match (3) Medium match (4) Low match (5) Very low match

18. To what extent do you think the results of the ITBS can be used for improving students' understanding of their specific strengths and weaknesses? (1) Not at all (2) To a mimimal extent (3) To some extent (4) To a great extent (5) To a very great extent

19. How useful are the results of the ITBS in helping parents better understand the strengths and limitations of their child? (1) Not at all useful (2) Minimally useful (3) Useful to some extent (4) Useful to a great extent (5) Useful to a very great extent

20. How well informed do you consider yourself to be about the ITBS? (1) Not informed (2) Minimally informed (3) Informed (4) Well informed (5) Extremely well informed

21. How would you rank the overall quality of the ITBS as compared to other standardized tests of its type? (1) One of the best (2) Above average (3) About the same as others (4) Below average (5) One of the worst

DIRECTIONS: This next set of multiple choice questions assesses your knowledge about the ITBS. There is one best answer for each. Please put the number of your answer in the right hand blank. If you are not sure of an answer, take a guess.

22. In attempting to determine whether or not the Iowa Tests of Basic Skills is appropriate for your school system, what is the most important issue to consider?

    (1) Does the test battery have sufficiently high reliability?
    (2) Do the test items of the battery correspond to the content of instruction in your school system?
    (3) Is the test battery based on a thorough survey of teaching practices over the whole country?
    (4) Is the student population upon which the test norms are based comparable to the student population of your school system?

    (2)

23. Which of the following greatly adds to the reliability of the ITBS results?

    (1) The homogeneity of the group tested.
    (2) The number of types of items on the tests.
    (3) The number of persons in the norming population.
    (4) The length of the test battery.

    (4)

24. The most serious criticism of the ITBS involves the
    (1) Unwise uses made of test results.
    (2) Inappropriateness of this test in measuring what is being taught in schools today.
    (3) Inappropriateness of comparing the scores of urban students to those of rural students.
    (4) The relatively low level of accuracy of test procedures.

    (1)

25. Skills analysis is <u>least</u> useful for
    (1) Planning instruction for groups of students.
    (2) Identifying individual strengths and weaknesses.
    (3) Measuring individual growth from year to year.
    (4) Identifying general class wide strengths and weaknesses.

    (3)

26. Which of the following is the biggest problem in interpreting the subskills of the ITBS:

    (1) The interpretation process is confusing for many students.
    (2) The low achieving students are not able to identify any strong areas.
    (3) The interpretation usually doesn't provide useful information about average students.
    (4) The differences between subskills can be over-emphasized.

    (4)

27. When doing a skills interpretation of the ITBS, it is most appropriate that the results be viewed as:

(1) Valid Measures of ability.
(2) Accurate measures of a student's progress in the subskills.
(3) Tentative indicators of strength and weakness.
(4) Definite guides to remediating weaknesses and capitalizing on strengths. (3)

DIRECTIONS: There is a best answer for each of the following True/False items. Please use "1" for TRUE and "2" for FALSE. Again, if you are not sure, please guess.

1 = TRUE 2 = FALSE

28. The ITBS show students' achievement in some school subjects that are important for future school success. (1)

29. A good use of the ITBS is to give grades at the end of each quarter or semester. (2)

30. The reading test of the ITBS measures three kinds of understanding: Facts, Inferences, and Generalizations. (1)

31. The ITBS measure all of the skills most students are taught in school. (2)

32. If a student misses most or all of the questions about some skill tested, it means that she or he does not know anything about that skill. (2)

33. If a student answers all the questions about a skill correctly, it means he/she has mastered that skill. (2)

34. One of the main purposes of the ITBS is to help students understand what their strengths and weaknesses are. (1)

35. The questions for each skill tested on the ITBS are all of about the same difficulty. (2)

APPENDIX B: STUDENT SURVEY

DIRECTIONS: The following ten questions ask for your opinions about tests. There are no right or wrong answers. Please answer the questions using the purple Standard Answer Sheet by filling in the space below the appropriate letter. If you have any questions, raise your hand.

1. How well do you think you did on the Iowa Tests of Basic Skills this year?

   A. Quite high
   B. Above average
   C. Average
   D. Below average
   E. Quite low

2. In general, how do you feel about tests that teachers make up?

   A. I really like them
   B. I like them
   C. I don't care one way or the other
   D. I hate them
   E. I really hate them

3. In general, how do you feel about the Iowa Tests of Basic Skills?

   A. I really like them
   B. I like them
   C. I don't care one way or the other
   D. I hate them
   E. I really hate them

4. How hard do you think tests like the Iowa Tests of Basic Skills are?

   A. Very hard
   B. Hard
   C. Medium
   D. Easy
   E. Very easy

5. How hard are the tests your teacher makes up?

   A. Very hard
   B. Hard
   C. Medium
   D. Easy
   E. Very easy

6. How nervous do you feel before you take a test that your teacher made up?

   A. Extremely nervous
   B. Very nervous
   C. Nervous
   D. Just a little nervous
   E. Not at all nervous

7. How nervous do you feel before you take a test like the Iowa Tests of Basic Skills?

   A. Extremely nervous
   B. Very nervous
   C. Nervous
   D. Just a little nervous
   E. Not at all nervous

8. How many questions on the Iowa Tests of Basic Skills cover things you have studied in school?

   A. All of them
   B. Most of them
   C. Some of them
   D. Only a few of them
   E. None of them

9. How much do you think you know about the tests on the Iowa Tests of Basic Skills?

   A. A lot
   B. Quite a bit
   C. A little
   D. Not much at all
   E. Nothing

10. How useful are the Iowa Tests of Basic Skills results to you?

   A. Extremely useful
   B. Very useful
   C. Useful
   D. Not useful
   E. Not at all useful

DIRECTIONS: The next set of questions tests your knowledge about the Iowa Tests of Basic Skills. They do have a right or wrong answer. The sentences below are either true or false. If you think a sentence is true mark an "A" on your answer sheet. If you think it is false mark a "B". If you are not sure, take a guess.

A = TRUE B = FALSE

11. Scores from the Iowa Tests of Basic Skills are most often used to find out what subjects, like science or reading or math, students are interested in.

12. The Iowa Tests of Basic Skills show how well students do in some school subjects that are important for future school success.

13. The Iowa Tests of Basic Skills tell how well students work together in groups.

14. The Iowa Tests of Basic Skills tell how students feel about the school subjects they study.

15 A good use of the Iowa Tests of Basic Skills is to give grades at the end of each quarter or semester.

16. Scores from the Iowa Tests of Basic Skills should almost always be used to help in making plans for students' future study.

17. The Iowa Tests of Basic Skills covers skills in math and reading but NOT language.

18. The reading test of the Iowa Tests of Basic Skills measure three kinds of understanding: Facts, Inferences, and Generalizations.

19. The Iowa Tests of Basic Skills test all of the skills most students are taught in school.

20. If a student misses most or all of the questions about some skill tested, it means that she or he does not know anything about that skill.

21. If a student answers all the questions about a skill correctly, it means he/she knows all the important things about that skill.

22. One of the main purposes of the Iowa Tests of Basic Skills is to help students understand what their strengths and weaknessess are.

23. A student should always do his/her best to answer the questions on the test in order for the test to be most useful.

24. The questions for each skill tested on the Iowa Tests of Basic Skills are all of about the same difficulty.

Key:	11. F	14. F	17. F	20. F	23. T
	12. T	15. F	18. T	21. F	24. F
	13. F	16. T	19. F	22. T	