

1-1-1992

Elementary school teachers' opinions regarding the purposes and interpretation of score reports from standardized achievement test batteries.

Edward J. Murphy

University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Murphy, Edward J., "Elementary school teachers' opinions regarding the purposes and interpretation of score reports from standardized achievement test batteries." (1992). *Doctoral Dissertations 1896 - February 2014*. 4902.

https://scholarworks.umass.edu/dissertations_1/4902

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST



312066013576613

ELEMENTARY SCHOOL TEACHERS' OPINIONS REGARDING
THE PURPOSES AND INTERPRETATION OF SCORE REPORTS FROM
STANDARDIZED ACHIEVEMENT TEST BATTERIES

A Dissertation Presented

by

EDWARD J. MURPHY

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 1992

School of Education

© Copyright by Edward J. Murphy 1992

All Rights Reserved

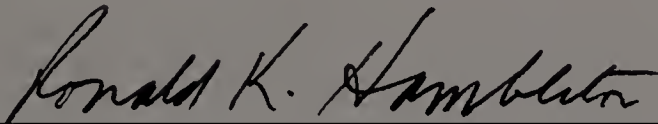
ELEMENTARY SCHOOL TEACHERS' OPINIONS REGARDING
THE PURPOSES AND INTERPRETATION OF SCORE REPORTS FROM
STANDARDIZED ACHIEVEMENT TEST BATTERIES

A Dissertation Presented

by

EDWARD J. MURPHY

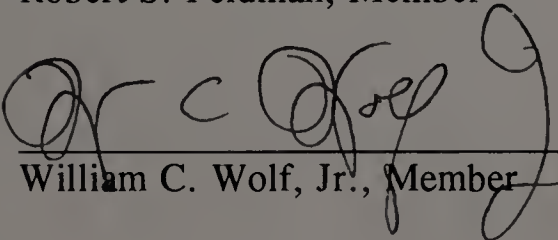
Approved as to style and content by:



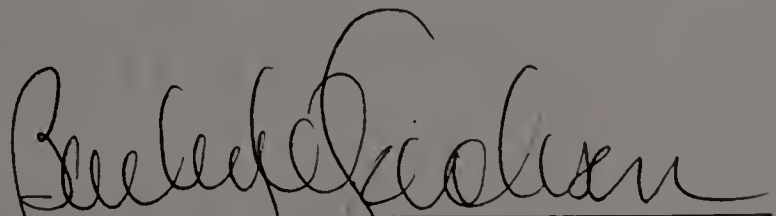
Ronald K. Hambleton, Chair



Robert S. Feldman, Member



William C. Wolf, Jr., Member



Bailey W. Jackson, Dean
School of Education

ACKNOWLEDGMENTS

The author wishes to acknowledge with gratitude the conscientious guidance of the members of his doctoral advisory committee, who provided many helpful suggestions for improving this study. Their advice and support were always considerate, aimed at making the study a more substantial contribution to educational research and at helping the author learn. In addition, a debt of gratitude is owed to the many colleagues and friends at National Evaluation Systems, Inc., who contributed time, resources, and talents to the conduct of this study and the preparation of this dissertation. The author would also like to express his appreciation to professional colleagues at the Texas Education Agency and the Illinois State Board of Education. Without their assistance and cooperation the study could not have been conducted effectively. Finally, to the "survey liaisons" who found the cooperating teachers to respond to the questionnaire and managed the distribution and collection of the instruments, and to those teachers who took the personal time to respond to an arduous questionnaire, the author extends his most sincere thanks.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS	iv
ABSTRACT.....	v
LIST OF TABLES.....	xi
Chapter	
1. INTRODUCTION.....	1
Background	1
Traditional Support for Standardized Tests	1
Criticisms of Standardized Tests	2
Purposes of Standardized Tests	3
Focus of This Study: Tests and Users	4
The Importance of the Classroom Teacher.....	4
Statement of the Problem	5
Purposes of the Study.....	8
Broad Summary of Method.....	8
Educational Importance of the Study	9
2. REVIEW OF THE LITERATURE	11
Introduction.....	11
Support for Standardized Testing	11
Testing Purposes.....	13
Publishers' Stated Purposes	18
The Role of the Teacher.....	20
Teacher Knowledge of Testing Issues	22
Proposed Solutions	25
The Reporting of Test Information.....	28
The Content of Reports	28
Language and Format of Score Reports	36
Conclusion	40
3. RESEARCH METHODOLOGY	41
Introduction.....	41
Overview of the Study Design	41

State of Assignment	86
Grade Level	86
Assignment Setting	88
Population of Municipality of Assignment	89
Teaching Experience	89
Training in Testing and Measurement	89
Usefulness of Testing Preparation	90
Frequency of Use of Testing Knowledge	90
Gender	91
Ethnicity	91
SATB Familiarity and Use	92
 The Section II Purpose Questions	 93
Form A/B	93
Mean Ratings on Purpose Questions	94
Differences in the Numerical (Report A) and Graphical (Report B)	
Ratings	96
Relationships between Independent Variables and the Section II	
Purpose Questions	99
Form N/S	101
Mean Ratings on Purpose Questions	102
Differences in the Narrative (Report N) and Numerical/Pictorial	
(Report S) Ratings	105
Relationships between Independent Variables and the Section II	
Purpose Questions	106
 The Section II Interpretive Questions	 110
Form A/B	111
Mean Ratings	112
Response Distribution	113
Comparison Across Report Formats	113
Relationships between Independent Variables and the Interpretive	
Questions	115
Form N/S	119
Mean Ratings	120
Response Distribution	121
Comparison Across Report Formats	121
Relationships between Independent Variables and the Interpretive	
Questions	123
 Open-Ended Comments	 128

5. CONCLUSIONS AND RECOMMENDATIONS131
 General Purposes131
 Score Report-Related Purposes and Preferences.....133
 Interpretive Questions135
 Recommendations137

APPENDICES

A. QUESTIONNAIRE FORM A140
B. QUESTIONNAIRE FORM S151
C. SUPPORT MATERIALS SENT TO SURVEY LIAISONS162
D. COVER MATERIALS TO RESPONDENTS169
E. LETTER TO MASSACHUSETTS RESPONDENTS173

BIBLIOGRAPHY175

LIST OF TABLES

Table	Page
3.1 Survey Packets Sent to Texas	64
3.2 Survey Packets Sent to Illinois.....	66
3.3 Survey Packets Sent to Massachusetts	67
4.1 Questionnaires Returned, by State	74
4.2 Respondents' Grade Assignments	74
4.3 Population of Municipality of Assignment, by State	75
4.4 Years of Teaching Experience, by State	76
4.5 Number of Courses and Workshops and Percent of Respondents in Each Category.....	77
4.6 Usefulness of Preparation in Testing and Frequency of Use of Testing Knowledge	77
4.7 Gender and Ethnicity, by State	78
4.8 Texas (1992) and Illinois (1990) Teachers' Gender and Ethnicity (%): Respondent Sample and Public School Teaching Population (K – 8)	79
4.9 SATB Familiarity and Use	80
4.10 Purposes of SATBs: Number of Respondents, Mean Ratings, and Standard Deviations	82
4.11 Purposes of SATBs: Distribution (Number and Percent) of Responses.....	83
4.12 Mean Ratings for Section I Questions, by State	87
4.13 Mean Ratings of Section I Questions with Significant ($p < 0.05$) Differences by Ethnicity	92
4.14 Section II Purpose Questions for Score Reports A and B: Number of Respondents, Mean Ratings, and Standard Deviations.....	95
4.15 Purposes of Score Report A: Distribution (Number and Percent) of Responses	97
4.16 Purposes of Score Report B: Distribution (Number and Percent) of Responses	98

4.17	Section II Purpose Questions for Score Reports A and B: Mean Ratings by State	100
4.18	Section II Purpose Questions for Score Reports N and S: Number of Respondents, Mean Ratings, and Standard Deviations	103
4.19	Purposes of Score Report N: Distribution (Number and Percent) of Responses	105
4.20	Purposes of Score Report S: Distribution (Number and Percent) of Responses	106
4.21	Section II Purpose Questions for Sample Score Reports N and S: Mean Ratings by State.....	108
4.22	Section II Purpose Questions for Sample Score Reports N and S: Mean Ratings by Ethnicity	109
4.23	Section II Interpretive Questions for Score Reports A and B: Mean Respondent and Local Psychometric Group Ratings for Related Questions.....	112
4.24	Score Report A and B Interpretive Questions: Distribution (Number and Percent) of Responses	114
4.25	Section II Interpretive Questions for Score Reports A and B: Mean Ratings by State.....	116
4.26	Section II Interpretive Questions for Score Reports A and B: Mean Ratings by Ethnicity	117
4.27	Section II Interpretive Questions for Score Reports N and S: Mean Respondent and Local Psychometric Group Ratings for Related Questions.....	120
4.28	Score Report N and S Interpretive Questions: Distribution (Number and Percent) of Responses	122
4.29	Section II Interpretive Questions for Score Reports N and S: Mean Ratings by State.....	124
4.30	Section II Interpretive Questions for Score Reports N and S: Mean Ratings by Ethnicity	125

CHAPTER 1

INTRODUCTION

Background

The use of standardized norm-referenced achievement tests in schools in the United States is prevalent, especially at the elementary education level. Every year, millions of students in U.S. public and private schools are administered one or more tests, often in the form of batteries covering several subject areas, designed to compare their achievement levels with those of students at similar grade levels across the country, or more precisely, with a norming sample of students who were administered a version of the tests while they were at the same grade levels during a previous year. The amount of money spent on these tests is considerable, and equally or more important, the amount of school time spent on them is large. Time and money are expended not only on the selection, purchase, and administration of the tests, but also on preparing students to take them, scoring them, interpreting them at various levels of the school population, and discussing them among their many audiences.

Traditional Support for Standardized Tests

Standardized tests are prevalent in the schools because they have traditionally been popular among most of their audiences [Rudman, 1977]. Officials concerned with the effective and efficient operation of public schooling as a national, state, or local issue have been interested in these sorts of tests for their promise of imposing some form of accountability and comparability (either nationally, internationally, or locally) on the efforts of those who operate and work in the schools [Mehrens & Lehmann, 1987]. Other members of the more general public (e.g., business people, taxpayers, parents) have also supported the administration of tests of this kind [Resnick, 1981]. The traditional support for testing, despite the controversy that has always surrounded this endeavor, has usually

been based on the perceived impartiality and objectivity of the tests, which contrasts with the apparent lack of objectivity of other forms of student assessment (e.g., classroom testing, teacher grading). The perception has appeared to be not that teachers are widely prone to arbitrary or excessively generous grading and evaluation practices, but that standardized tests impose a salutary layer of external verification on a system that is otherwise relatively unsupervised and autonomous [Resnick, 1982].

Even teachers have traditionally supported standardized testing, at least for some purposes and within clear constraints of timing and interpretation. Among teachers, the perception has apparently been that standardized tests can sometimes provide useful information to complement the information that they routinely gather less formally through a variety of other means [Rudman, 1977; Salmon-Cox, 1981]. The information has generally been expected to be confirmatory rather than surprising, although unexpected test scores may have induced teachers to reevaluate their perceptions of the students who attained them [Gardner, 1982; Salmon-Cox, 1981].

Criticisms of Standardized Tests

The traditional support for standardized testing in the schools does not imply that such testing is uncontroversial. In fact, a great deal of controversy surrounds this issue, and standardized tests are often the target of heated opposition involving many objections. Standardized tests are accused of oversimplifying the assessment of complex human attributes [Ravitch, 1983-84] and of labeling children with virtually indelible marks that become not merely announcements but prophecies [Kolstol, 1967; Mehrens, 1967]. Teachers and administrators are indicted for relying too heavily on tests for grouping, classifying, and determining the content of instruction for students [Airasian, 1980; Madaus, 1985]. As a general phenomenon, standardized tests are held responsible for causing instruction to become narrow, minimal, and reductionist because of the tests' allegedly narrow, minimal, and reductionist conceptualizations of human learning [Madaus,

1988; Ravitch, 1983-84]. Moreover, testing is said to dominate classrooms in terms of the time it occupies in an already busy and overambitious schedule, to foster competition among students, and to encourage simplistic comparisons (of classrooms, of students, of teachers, of teaching styles, of resources and materials) by school and district administrators, parents, legislators, and the general public [e.g., Neill & Medina, 1989; Rayborn, 1989]. Finally, standardized tests are often called biased by their critics: biased against cultural, ethnic, racial, and learning style minorities and against females [discussion, for example, in Ebel, 1976; Gardner, 1978; Jackson, 1975; Williams, 1971].

Both the traditional support and the ongoing controversy underscore the importance of standardized achievement tests as a phenomenon in the educational culture of the United States. And given this importance, fair and reasonable questions can be raised about the purposes, advantages and disadvantages, uses and misuses, and justification of standardized testing as an educational activity. Such questions often proceed from the assumption that anything as resource-costly as standardized tests in so resource-poor an environment as the United States public school must serve a truly useful purpose if it is to be justified.

Purposes of Standardized Tests

Standardized tests have been claimed and expected to perform a range of purposes in support of American education. The purposes include helping the teacher determine the level of performance achieved by individual students in the subject areas covered by the test; providing diagnostic information to the teacher about individual students' strengths, weaknesses, and needs; providing similar information on groups of students to help the teacher form temporary study or work groups; helping the teacher (or the school) make placement decisions (e.g., into grade levels, into basal readers, into special classes); informing the teacher about the effectiveness of his or her instruction on particular topics; giving students insight into their own areas of strength and weakness; offering unexpected insights to teachers into the talents, interests, and achievements of students; helping

teachers, curriculum supervisors, and other school personnel evaluate curricula, materials, resources, and approaches at the classroom level and design modifications to address problem areas or emulate unusually successful strategies; providing information to administrators for evaluating the effectiveness of teachers in meeting school expectations; helping building-level or district-level school officials target approaches that have proven effective and ineffective in different classrooms within the school building or across buildings or years; and at the district level, assisting superintendents and school boards in their evaluations of the effectiveness of schools, principals, teachers, instructional supervisors, or instructional approaches and materials.

Focus of This Study: Tests and Users

Not all of these purposes are equally accepted by educators and not all are equally feasible for all kinds of standardized achievement tests. The type of standardized achievement test that will be the focus of this study is what Mehrens and Lehmann [1987] refer to as the *standardized achievement test survey battery*. This term signifies a comprehensive battery of tests in several basic skills subject areas (e.g., mathematics, language arts, social studies) normed on the same sample of students; the most common examples of the survey battery are the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Test of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), and the Stanford Achievement Test (SAT). There are several intended audiences for the results derived from these test batteries; the audience that will be the focus of this study is the regular elementary-level teacher.

The Importance of the Classroom Teacher

It should be clear from the list of purposes of standardized achievement testing above that the classroom teacher has an important role with regard to testing. At least at the elementary school level, the most frequent potential user of survey battery results is

probably the teacher. Whether the results are to be used to individualize instructional approaches, form temporary groups to work on weaknesses or capitalize on strengths, confirm hypotheses about the skills of particular students, or adjust classroom strategies to teach a concept or group of concepts differently, the teacher is the most appropriate audience. In most cases it is the teacher who not only administers the test battery, but also receives the scores of the students and is expected to interpret and use them. If the purposes and uses of these batteries are important, then the role of the teacher is crucial. And if the teacher must play a crucial role, he or she must be equipped to play it effectively or the testing enterprise cannot have positive results for the students and the school.

In short, the teacher is often the key to the most important aspect of the testing program: the use of test results to effect positive change.

Statement of the Problem

According to the preceding analysis, the value of standardized achievement testing, in relation to its costs in time and money for the schools, resides in the effectiveness of these tests in achieving significant purposes essential to schooling; and a large responsibility for attaining these purposes at the elementary level is held by the classroom teacher. It then becomes important to be assured that the teacher can indeed use the tests to meet these purposes. This is where the problem emerges. There is substantial evidence that the regular classroom teacher at the elementary level may not be adequately prepared to deal with and use the wealth of information that can be derived from standardized achievement test survey batteries, or at least not to use it in its current form, and may in fact be overwhelmed by the vast quantity of this information and other related information provided by the publishers of such batteries.

All of the survey batteries mentioned above provide a great deal of information in the form of test results, and all of them offer helpful ways of breaking down, aggregating, and arraying these results for a variety of purposes. In addition, all of the batteries offer

volumes of interpretive guidelines designed to help the teacher understand such concepts as percentile, grade equivalent, median, standard deviation, confidence band, and so on.

Nevertheless, three factors related to the preparation of the teacher for his or her job and to the nature and performance of that job itself militate against the teacher's effective use of this information in the form in which it is presented by test publishers.

First, the preparation of most teachers for their jobs involves at most one or two courses in which testing issues, terms, and concepts may be discussed; even if such issues are discussed, there is little control over the nature of the instruction and little assurance that the teachers-to-be understand it [e.g., Durost, 1959; Mayo, 1959; Noll, 1955; Roeder, 1972]. There is also evidence that education majors shy away from courses on tests and measurements because of their perception that such courses are dry and difficult [Durost, 1959; Gullickson, 1986]. Teachers' opportunity and will to learn testing concepts during preservice education may therefore be quite limited [Hagen & Lindberg, 1963].

The second factor, related to the time when the teacher is actually performing the job, also constrains the teacher's effectiveness as a user of test results. The job of the teacher is time-consuming: he or she must plan instruction for every student, evaluate student progress, attend numerous meetings with colleagues and parents, and deal with emergencies—all in addition to actually teaching. In this welter of responsibilities it may be too much to expect the teacher to spend time reading the helpful and instructive information about the interpretation of test results that the publishers of the survey batteries provide.

The third factor relates to the disparate cultures that surround the enterprises of teaching and testing [Tittle, 1989]. Teachers inhabit a culture in which instructional decisions have to be made instantaneously and somewhat intuitively, based on a variety of tangible and intangible events, characteristics, and nuances within the classroom. The contexts within which they interpret test score information may be entirely different from the ones imagined by equally conscientious but differently enculturated psychometricians. The languages that many teachers and psychometricians speak may be different not only in

the nature of their technical vocabularies, but also in the cultural understandings and givens that imbue them unconsciously. The meanings that teachers construct from score reports may be quite surprising to the authors of those score reports.

The consequence is predictable: There is considerable evidence that teachers lack thorough understanding of the essential testing concepts upon which the psychometrically accurate interpretation and use of test results depend [e.g., Campbell, 1981; Culyer, 1982; Hills, 1991].

Most solutions to this problem have focused on the preparation of the teacher for the complex job of test results interpretation. Calls for increasing the number of courses in testing and measurement required of teacher education students at the undergraduate or graduate level are frequent and chronic. Similarly chronic are calls for inservice education in such concepts for practicing teachers. The problem with both solutions is time: too little time in the teacher preparation curriculum for additional testing courses, too little time in the teacher's inservice training schedule for substantial testing courses, and too little time in most inservice courses to cover the topics that most psychometricians regard as essential.

Furthermore, as far as preservice training is concerned, a major educational movement of the day actually runs counter (probably unintentionally) to the desire to increase teacher candidates' exposure to testing concepts. The trend in teacher education is away from "educational foundations and methods" courses and toward liberal arts and sciences courses. This trend is accompanied by a related trend toward providing alternative routes to certification (e.g., not through the schools of education) to candidates with general collegiate preparation in any number of disciplines. In the light of such tendencies, the hope for increasing the amount of time future teachers spend on measurement concepts seems dim.

Perhaps, then, it is advisable to consider reversing the terms of the proposed solution, at least in part. Instead of trying to shape teachers into surrogates for psychometricians by equipping them with greatly increased amounts of technical information and understanding,

it might be more feasible to tailor the test-related information that they do receive to their own capabilities and constraints—and to their own pedagogical wisdom, needs, and culture—not as a substitute for preservice and inservice education, but as a complement to those approaches. And perhaps one good way to learn just what teachers do and do not find useful in terms of test results, and can and cannot understand in terms of test interpretations, is to ask them. This study is an attempt to do that.

Purposes of the Study

The overall purposes of this study were:

1. to explore teachers' opinions regarding the purposes for which the major standardized achievement test survey battery score reports are useful;
2. to explore teachers' opinions regarding the content and format of typical score reports from such batteries; and
3. to explore the psychometric appropriateness of teachers' interpretations of test results presented in varying ways.

Broad Summary of Method

This study attempted to achieve these purposes through the preparation, administration, and interpretation of a survey of the opinions of more than 650 elementary school teachers from Massachusetts, Illinois, and Texas. The teachers were asked to respond to demographic background questions; opinion questions (via a Likert-type scale) regarding frequently stated purposes of standardized achievement test survey batteries; opinion questions based on sample, hypothetical score reports containing different contents and formats; and knowledge questions regarding the interpretation of results presented in different conditions of content and format. The results from the teachers' completed

questionnaires were tabulated and analyzed to seek answers to the research questions posed by the study.

It has been the intent of this study to be exploratory. Efforts were made to draw voluntary samples of public elementary school teachers from a variety of types of schools, demographic backgrounds, and geographic regions within a varied group of states. The teachers selected to participate were expected to have personal familiarity with at least one of the major standardized achievement test batteries. To support the interpretation of survey results inferentially with respect to the larger population of public elementary school teachers in the United States, the actual sample of respondents has been characterized in terms of demographic factors that may be used to reveal their typicality or lack of typicality in terms of that larger population.

It is the researcher's hope that others can build upon the information gained through this study to continue exploration of an important educational issue.

Educational Importance of the Study

Testing is established as a significant activity undertaken by schools in this country, and the prevalence of standardized achievement test survey batteries is unlikely to diminish soon. Because such tests are used for a number of important purposes, they affect millions of students annually in highly significant ways, including helping to determine the content to which students are exposed, the settings in which students learn, the colleagues with whom students spend large amounts of their time, and the resources and strategies that will be used to instruct them. Beyond the confines of strict instructional decisions, test results may also affect the opinions that teachers will construct regarding their students' abilities, students' own self-concepts, their parents' opinions regarding their abilities, and even the students' educational and career choices.

The interpretation and use of test results at the elementary level is typically almost entirely in the hands of the classroom teacher. This is a highly important responsibility. It

has been demonstrated that classroom teachers' background and understanding of test results may not be strong. Proposed solutions to this problem that focus on increasing teachers' knowledge of testing and measurement issues through preservice or inservice education may be unfeasible.

This study is intended to contribute information toward a solution of the problem of teachers' need for testing knowledge by helping to increase understanding of teachers' preferences regarding the use and provision of test results, and of teachers' ability to understand results presented in different ways. If this study, by helping to focus test results reports on issues of genuine concern to teachers and by helping to guide the formatting of reports in ways that teachers prefer and can understand, contributes to bridging the gap between the information that is available to teachers through tests and teachers' ability to grasp that information, it will have made a contribution to solving an important educational issue. As the review of the educational literature that follows in Chapter 2 shows, not very much work has been reported that focuses on ascertaining teachers' preferences regarding the reporting and use of results of standardized achievement tests or on teachers' ability to comprehend various formats through which test results are commonly reported.

CHAPTER 2 REVIEW OF THE LITERATURE

Introduction

The administration of standardized tests in U.S. schools is a common phenomenon. One recent estimate [cited in Pikulski, 1989] reported that approximately 105 million standardized tests were administered every year to the 39.8 million students in classrooms in the United States in the mid-1980s, or about 2.6 tests per student per year. Another estimate [cited in Anderson, 1982] reported that, during the course of his or her 13 years of schooling, the average U.S. student will have taken 6 to 12 comprehensive test batteries. Looking at this issue in financial terms, Whitehead and Santee [1987] estimated that standardized tests were costing the schools (i.e., taxpayers) \$40 million per year to administer, a figure that accords well with Resnick's [1981] estimate of \$42 million per year in 1976.

It is not clear whether the current trend is toward more or less use of standardized achievement tests, or remains about the same as it has been. Resnick [1981] compared the \$42 million per year figure for 1976 with the corresponding dollar figure for 1948. Adjusting for inflation, the 1976 figure equates to about \$24 million; the amount spent on such tests in 1948 was \$7 million. On the other side of the coin, Davis [1962] reported that in 1958 about 122 million tests were administered; compared with the 105 million figure for the 1980s reported above, this may reflect a decline in test use to match demographic trends over the same time period, or, perhaps, an incremental change in attitude toward this form of testing.

Support for Standardized Testing

Whatever the trend for the use of such tests, it seems reasonable to conclude that the standardized testing phenomenon will not disappear in the near future. This conclusion

appears to be reasonable because, despite recent complaints that there is too much testing in schools [e.g., Jacobs, 1988], the fact remains that such complaints would have to overcome a great deal of traditional support for the use of standardized tests in the schools. Probably the most vocal support has come from the general public (whose opinions of course affect those of legislators and policymakers). Many fairly recent reports refer to the utility that the public finds in such tests as a way to place an external quality check on the relatively unfettered enterprise of U.S. public education [Mehrens & Lehmann, 1987; Dreher & Singer, 1985; Resnick, 1981, 1982; Rudman, 1977]. At least into the 1980s, the tradeoff in the mind of the public has appeared to be to allow teachers and other school officials a relatively free hand in the classroom in exchange for imposing the comparatively benign and impartial monitor of the standardized test. Nor has the support of standardized tests been merely a cost-value matter for taxpayers; parents, a very important sector of the general public with an even greater stake in public education than money, have also traditionally supported tests in the schools [Anderson, 1981; Dreher & Singer, 1985].

It is perhaps unremarkable that support for standardized testing has been found among school board members [Boegli et al., 1977], given their emphasis on outcomes, performance, and accountability, but it is more surprising to find documented support, even fairly recently, for this form of testing among teachers, who are often depicted as opposing any assessments that they have not devised themselves. However, Salmon-Cox [1981] encountered support for, and use of, standardized tests among teachers, as did Cummings and Stinard [1983] and Gullickson [1984]. These findings corroborate the generally positive attitudes of teachers (and other school populations) toward standardized achievement testing that had been found by Goslin [1967] in his important earlier study of this issue. Not only did elementary school teachers in the Goslin study report that standardized achievement test batteries were useful for many purposes, but they also felt that "about the right number" of such tests were given in their schools each year.

There is some recent evidence of the emergence of a more skeptical and reserved attitude toward standardized testing among elementary school principals. A survey of more than 800 principals sponsored by the National Association of Elementary School Principals (NAESP) was briefly reported in the March 11, 1992, edition of *Education Week* (page 1 of a special advertising supplement called "Conventions in Print"). The survey reportedly found that elementary school principals were "fairly evenly divided on the question of whether standardized tests should be used in elementary schools at all." It was also reported that the respondents were "overwhelmingly opposed" to the use of such tests with young children, were skeptical of standardized tests as measures of basic skills, and doubted the ability of the general public to interpret test scores accurately.

Testing Purposes

Support for standardized achievement tests is dependent on the useful purposes that such tests serve or are perceived to serve. Traxler [1960] compiled a catalog of ten potential uses of the results of large-scale testing programs: to delineate a curricular starting point for classes within a school by comparing the characteristics of the school's population with those of a national sample; to compare individual students' achievement and aptitude levels; to help teachers know the level of ability of each of their classes; to help teachers know the level of ability of each of their students; to help teachers diagnose student needs; to help teachers and students discover special abilities; to provide teachers and students insight into students' interests; to provide educational and vocational guidance; to help students with adjustment problems; and to support educational research.

Traxler's list, though long, does not focus on the ultimate uses to which standardized test information might be put. In a shorter catalog developed 14 years earlier, Campbell [1946] was more specific about such ultimate uses, the area in which most controversy about test use exists. Campbell, prefiguring some later criticisms of standardized testing, stated that the testing movement had done good work but worried that it was coming under

a cloud because some enthusiasts had oversold the benefits of testing. The result (in 1946) was that too many schools used test results merely to classify students. According to Campbell, this was an underuse because tests could help teachers help students. In support of this thesis, he cited several valid uses for test results, including the early detection of group problems with instruction and learning; diagnosis of student needs and provision of educational prescriptions to meet them; support for placement and promotion decisions; help in organizing remedial and special classes; and provision of information for research purposes.

In his 1967 study of educators' attitudes toward testing, Goslin encountered general support for several uses of standardized achievement test batteries in the elementary schools. In terms of frequency, achievement test batteries were the most often used type of standardized test in the elementary grades (37.1 percent of all tests reported were of this type; the next highest reported type of test was 24.6 percent for group-administered intelligence tests). In rank order of frequency of use, elementary principals reported using such batteries for diagnosing learning difficulties (78.8 percent of reported uses of this sort of test), homogeneous grouping (39.4 percent), counseling children (34.0 percent), evaluating the curriculum (33.1 percent), and counseling parents (30.4 percent). There were far fewer citations and less support for using such batteries for grading students (9.1 percent) or evaluating teachers (4.2 percent).

The latter two potential uses of standardized achievement test results are the subject of much criticism but the object of very little overt support even among test enthusiasts. Determining students' class grades solely, substantially, or even partially on the basis of the students' performance on a wide-ranging standardized achievement test battery is universally decried even in the literature that is generally supportive of testing [e.g., Linn, 1983; Mehrens & Lehmann, 1987; Rupley, 1973; Salmon-Cox, 1981] for clear reasons of probable lack of fit between the content of such tests and curricular content and because of the reliability problem inherent in the relatively small numbers of test items typically

composing the subtests of such batteries. The only possibly legitimate use of such test results in relation to grading, according to Mehrens and Lehmann [1987], is to help teachers evaluate their own grading practices (i.e., to help them assess whether they are generous or strict graders relative to other teachers).

As for the use of standardized test results in the evaluation of teachers by school or district administrators, this practice is neither widely attested nor supported in the literature; it is of course possible that actual practice may differ from reported practice in this matter. Mehrens and Lehmann [1987] flatly condemn the use of standardized achievement test scores as the sole measure of teacher effectiveness, although they do recommend that such scores be "used judiciously as *one of many* variables in teacher evaluation" [p. 307; emphasis in original].

As reported above, Goslin [1967] found few citations of this test use among elementary principals (4.2 percent of all reported uses). He also found that teachers generally disapproved of this practice, with 29.1 percent of the 86 responding elementary teachers agreeing that standardized achievement test scores should never be used to evaluate a teacher's effectiveness and an additional 59.3 percent agreeing that the practice is "a relatively poor way of evaluating a teacher's effectiveness." It is more surprising that as many as 11.6 percent of the teachers agreed that such a use was sometimes the best way of evaluating a teacher's effectiveness.

Reporting on the administrative perspective on a related, but higher-level, evaluative use of standardized tests, Sproull and Zubrow [1981] found little support in central (i.e., district) offices for the use of such test scores as school quality measures. In their study, central office administrators stated that such test results might be used at the central office level as indicators of problems to be worked on at the building level, but in fact expressed the belief that others, including principals and teachers, benefited from and used test results more than the central office. In a similar vein, Plumleigh [1977] argued that standardized

achievement test scores could be useful for raising questions for the schools to work on, but should not be used to pontificate about school failures or to applaud school successes.

As a general finding in the literature, there has been substantial support among teachers and other school officials, at least into the 1980s, for the use of standardized achievement test results for certain specific purposes. Salmon-Cox [1981] reported on a survey of teachers in which about half of the comments indicated that standardized test results were used as a supplement to or a confirmation of information teachers had already obtained from other means, about one-fifth of the comments indicated that such test results were used as a reflection on or a guide to instruction, and about one-fourth indicated that such results were used to confirm (not make) grouping and tracking decisions for students. A less expected finding relates to what teachers reported doing when a surprising score resulted from a standardized achievement test: generally, if a student scored less well than expected, the test score was discounted as an aberration, but if a student scored better than expected, the score was treated as a red flag indicating that the teacher ought to look more closely at the student in case the teacher had been missing a hidden ability.

In their study of teachers' opinions about standardized tests, Stetz and Beck [1981] presented to teachers eight possible uses of test results and found that more than half of the surveyed teachers reported using such tests for diagnosing student strengths and weaknesses (74 percent), measuring growth (66 percent), evaluating individual students (65 percent), and planning instruction (52 percent). The other potential uses suggested by the authors also garnered considerable support: class evaluation (45 percent), reporting to parents (42 percent), evaluation of teaching methods (37 percent), and reporting to students (24 percent). Ladd [1971] recommended that teachers use unorthodox methods of manipulating and analyzing standardized test data to derive useful diagnostic information (i.e., steps beyond those suggested by test publishers, such as analyzing student errors, observing students' test-taking patterns and patterns of performance, and plotting test score distributions to spot outliers), a recommendation that at once supports the information that

can be gleaned from standardized tests and undermines the utility of more orthodox methods of using such results that do not require special manipulations.

In a more skeptical vein, Linn [1983] reported the Stetz & Beck [1981] information mentioned above, but also cited other studies that indicated that teachers did not use information from tests to deal with student problems or problems with their own teaching, relying instead on their own judgments. He struck a theme that has been echoed elsewhere in the literature, particularly with reference to the use of standardized test information to improve instruction: "Although the instructional usefulness of present-day standardized tests may be debatable, not even the strongest advocate could be satisfied that the lofty stated purposes of test publishers are fully realized" [p. 182]. A similar skepticism about publishers' claims was expressed by Buros [1977]. Gardner [1982] also complained that the diagnostic usefulness of such tests might be overemphasized and overrated.

Other researchers have stressed that useful purposes can be served by standardized achievement test scores provided that they are used and interpreted in a broader context of teacher knowledge and experience relative to students. As Rupley [1973] put it: "Test scores should be interpreted in conjunction with the total child, his home life, prior performance on tests, prior teacher evaluations and the other myriad factors related to his total makeup" [p. 755]. And Leiter [1976] argued that standardized tests, designed to "cleans[e] the evaluation process" [p. 59] of unreliability by eliminating the bias of teachers' background knowledge, not only do not do so but also should not do so. In fact, Leiter's argument is that "it is through the use of background knowledge that the objectivity of the test is secured by rendering an otherwise truncated account of the student's capabilities into a rich and immediate context of tacitly and explicitly known matters" [p. 65]. In other words, test results, to be meaningful, must be interpreted in a context of what the teacher already knows about the student.

In summary, then, some possible uses of standardized test results may be regarded as generally discredited, including the grading of students and the evaluation of teachers, at

least without the concomitant use of other confirmatory information that is more reliable. Other possible uses of standardized test results seem to be more acceptable, especially when the results are interpreted within a context of teacher knowledge about students and are used tentatively to confirm or be confirmed by other information. These uses include helping the teacher identify student needs and possibly beneficial instructional interventions, helping the teacher form temporary study or work groups based on similar needs, providing unexpected insights into student strengths and interests, supporting self-evaluation by teachers and evaluation by them of instructional approaches and resources, and providing similar information relative to broader program evaluation for teachers and other school officials to use in considering curricular or resource changes. These purposes, while not entirely uncontroversial, have been more widely accepted as reasonable for standardized achievement test survey batteries, at least into the 1980s.

Publishers' Stated Purposes

It should not be surprising that the purposes of standardized achievement test survey batteries as stated by their publishers generally accord with the above-cited purposes. The publishers of these batteries are among the most sophisticated users of and contributors to the research on legitimate purposes for their instruments. While evidently wishing to make their products appealing to potential customers by stressing their utility, publishers are concurrently aware of the dangers of overstatement and exaggerated claims. The result is generally caution in stating the purposes for which their tests are appropriate, combined with some ambiguity to permit varying interpretations by customers.

For example, the Riverside Publishing Company is explicit in stating "some of the specific purposes which the *Iowa Tests [of Basic Skills]* were designed to serve:"

1. to determine the developmental level of each pupil in order to adapt materials and instructional procedures more precisely to individual needs and abilities;
2. to diagnose specific qualitative strengths and weaknesses in a pupil's educational development;

3. to indicate the extent to which individual pupils have the specific readiness skills and abilities needed to begin instruction or to proceed to the next step in a planned instructional sequence;
4. to provide information useful in making administrative decisions in grouping or programming to accommodate individual differences;
5. to diagnose strengths and weaknesses in group performance (class, building, or system) which have implications for change in curriculum or instructional procedures or emphasis;
6. to provide a behavioral model to show what is expected of each pupil and to provide feedback which will indicate progress toward suitable individual goals;
7. to report progress in learning the basic skills to parents in objective, meaningful terms.

[From *Teacher's Guide: Multilevel Battery Levels 9-14, ITBS Forms G/H*, Riverside Publishing Company/University of Iowa, 1986.]

With the exception of #6, which is not a commonly stated purpose of standardized tests, these purposes fit with the purposes that have been stated in the literature. It is noteworthy, however, that a certain vagueness in terminology and in agent for these purposes is built into the above description (e.g., Does #3 include only within-class "steps in a planned instructional sequence," or is grade-to-grade promotion also encompassed? Who makes the administrative decisions alluded to in #4? Who diagnoses in #5, and what implications for change are included?), perhaps in recognition of the controversy surrounding the uses of standardized test information.

Even less explicit than the ITBS purposes are those of CTB/McGraw-Hill's

Comprehensive Tests of Basic Skills (CTBS):

A well-planned comprehensive testing program built upon a high-quality test provides information that supports the decision-making process in many areas, including the following:

- Needs Assessment
- Instructional Program Planning
- Pupil Analysis
- Program Evaluation

- Curriculum Analysis
- Class Grouping
- Evaluation of Student Progress
- Administrative Decisions

[From *Comprehensive Test of Basic Skills, Fourth Edition, Test Interpretation Guidelines*, Monterey, CA: McGraw Hill, Inc., 1988.]

A reasonable interpretation of these purposes concludes that they are very much aligned with the purposes of standardized achievement test survey batteries outlined in the literature. The vagueness of description is most likely a recognition of the necessity to avoid outright overstatement.

The Role of the Teacher

For most of the legitimate purposes of standardized achievement tests the teacher is the most obvious agent. It is the teacher who must most often make or contribute to placement decisions, diagnose individual students' needs, diagnose class needs, form temporary groups for targeted instruction, devise or change instructional approaches, assess student progress and growth, and discuss progress and growth with students and parents. If these teacher responsibilities are truly to be facilitated and guided by the results of standardized test batteries, the teacher must understand how to interpret those results for those particular and different purposes. This responsibility is a heavy one for the teacher to bear.

Teachers' responsibilities relative to the interpretation and use of test results have long been recognized and insisted upon in the literature. In providing a primer of types of tests in use at the time, Clark [1957] stressed the importance of testing knowledge to all teachers. His refrain was echoed by Conant [1963] in his influential work on the education of American teachers and by Mayo [1964] in the very first issue of the *Journal of Educational Measurement*. Of 70 statements about testing and measurement issues

submitted to teachers, principals, professors, and testing and research specialists by Mayo, the ability to interpret achievement test scores ranked second in importance. Goslin's [1967] study of teachers and testing concluded that teachers needed increased training in testing issues in order to meet their essential responsibilities relating to test use.

In 1968, when the issue of sharing test results with students and parents was emerging, Pounds and Hawkins concluded that parents had a right to information on the tests taken by their children, and consequently that teachers and principals had a growing responsibility to understand test scores in such depth that they could make test results meaningful to parents, in terms parents could be expected to understand. Teachers' understanding was to encompass the meaning of test scores and the amount of confidence it was appropriate to place in them. Pounds and Hawkins recommended the use of graphic displays and charts to help parents understand their children's results. Rupley [1973] placed high expectations on teachers for the amount of understanding they should possess relative to testing terms and concepts, calling for a rich interpretation of test scores in the context of the whole child.

More recently, Anderson [1981] also addressed the issue of teacher knowledge in terms of parents' rights, calling upon teachers to accept the responsibility for knowing and explaining the types of tests used in the schools and the limitations and legitimate uses of each one. Popham and Hambleton [1990] pointed out that testing was becoming more and more important in schools and that teachers and administrators had to sink or swim "in this testing maelstrom" [p. 38]. Making the point that testing had changed significantly in recent years, the authors called upon teachers and administrators to accept the responsibility for learning about certain essential issues in educational measurement, including the interpretation and use of test scores.

In the more recent years, a growing emphasis has been placed on teachers' knowledge of testing concepts relative to classroom tests [e.g., Linn, 1990; Stiggins, 1991a, 1991b], but the interpretation of standardized tests will remain important as long as such tests are

administered. Both Linn and Stiggins include knowledge about standardized tests in their domains of essential teacher knowledge, as does Schafer [1991] in his recent article on the required skills of teachers.

Teacher Knowledge of Testing Issues

In general, the assessment of teachers' levels of understanding of testing issues has not matched the high levels of importance attached to that understanding. With few exceptions, teachers have been found to have significant gaps in their knowledge of testing issues. Rudman [1977] may have found teachers not only supportive of standardized achievement testing but knowledgeable about it as well, but his assessment was based on teachers' self-reports. Gullickson [1984] also found that teachers felt they were knowledgeable about testing, although they admitted that they had acquired their testing knowledge not during their preservice training, but on the job. Gullickson [1982] had already concluded based on an earlier study of rural educators in South Dakota that teachers needed additional education in testing issues, and in a later [1986] study involving teachers' and college educators' opinions of the importance of various testing issues, he found significant discrepancies in the amount of importance the two groups assigned to most of these issues.

In a study reported in 1987, Huebner found that a group of 45 regular education teachers used test information appropriately to make recommendations for the classification of hypothetical students as learning disabled, but this study stands out among many others, including reports by the same author, that tend to indict teachers for their inability to use test information accurately.

As long ago as 1929, Madsen reported that teachers were "hopelessly befuddled as to the details in giving, scoring and interpreting [standardized] tests." He found that 15 out of 43 teachers in his Educational Tests and Measurements class made 33 scoring errors in a class assignment that involved tallying the results of a standardized test. He also reported

on two other studies of teacher error in scoring, in one of which 77 percent of the papers had to be rescored because of scorer error. His recommendation was that teachers should receive inservice training. A similar report was made by Daggett [1934] a few years later, mostly pertaining to false statements made by teachers regarding intelligence tests.

The situation had not improved noticeably by 1958, when Phillips and Weathers reported based on a study they conducted that, of 5017 scorings of the Stanford test, 28 percent contained errors made by teachers. Similarly, Crook [1959] reported several typical interpretive misunderstandings by teachers in elementary schools. Ebel [1961], focusing mostly on classroom tests, cited a number of teacher errors in testing practice and called for more preservice and inservice training to remedy the situation.

Fredrickson and Marchie [1966] found that teachers were too trusting of test scores, even when healthy skepticism was called for; as a result, the authors recommended either the use of confidence bands, the cooperation of counselors in interpreting test results, or increased inservice training of teachers. Leiter's [1976] finding that teachers applied substantial background knowledge to the interpretation of test results, reported above, even though regarded by that author as salutary, should be mentioned here. Whether such an addition of subjective knowledge to test information is indeed beneficial is debatable; the fact that it was generally unintentional on the part of the teachers indicates a knowledge deficiency.

In more recent times, teachers' knowledge gaps continued to be the subject of scholarly attention. Campbell [1981] noted teacher misconceptions about testing, and underscored the importance of teachers as accurate communicators of test information to students and parents, especially in an era of open school records. Culyer [1982] also decried the confusion and misinformation about student achievement testing that he found to be common among teachers, principals, supervisors, and government administrators, even in a post-Privacy Act environment. Hills [1991] charged that teachers and administrators have significant gaps in their testing knowledge, a situation that he and many others find

unacceptable. Most recently, Impara, Divine, Bruce, Liverman, and Gay [1991] found that, while teachers' overall performance on a brief score report-based test was "not alarmingly low" [p. 17], the teachers displayed significant gaps in their knowledge and interpretations of score reports and that their performance was substantially aided by the presence of interpretive information during testing.

To examine the need for teacher training in score interpretation, Huebner [1988] conducted a study of teachers' decisions regarding special education placement recommendations for hypothetical students based on test information presented in a variety of ways. Fifty-one teachers (all but one of them regular classroom teachers) were asked to consider whether special education or regular education would be more appropriate for given students, represented only by test scores. The scores were presented as grade equivalents, percentile ranks, and deviation IQs; equated scores were presented for each student. The teachers were more likely to recommend (inappropriate) special education placements for students on the basis of percentile ranks than on the basis of (equated) grade equivalent scores or deviation IQs. Huebner concluded that the literature on overidentification of students as learning disabled had neglected an important variable: the type of score used to portray the test information. She recommended caution in using percentile ranks, since they appeared to be easily misinterpreted.

It should be noted also that, in a replication of this study with school psychologists instead of regular classroom teachers as the score interpreters [Huebner, 1989], the same pattern of results emerged, despite the psychologists' greater amount of training in testing issues. Furthermore, school psychologists were the subject of another study, this one by Ross [1990]. The findings were similarly discouraging: the psychologists made inconsistent choices in evaluating discrepancy scores, generally failing, despite their training, to use significance tests, and also made other misinterpretations of test scores. A related finding was also reported by Peckens and Bennett in 1968 for a sample of 25 high school counselors. The counselors made misstatements about intelligence tests during taped

interpretation sessions, despite their specific training in the proper interpretation of such tests. This sort of pattern was enough for Goldman in 1972 to call tests and counseling "the marriage that failed." Perhaps such findings should lead psychometricians to question whether additional preservice training in testing issues for all teachers will prove adequate, without other measures being implemented as well, to eradicate teachers' deficiencies in testing knowledge.

Proposed Solutions

By far the most commonly suggested remedy for the problem of the teacher knowledge gap relative to testing is additional training during the preservice preparation for the teaching profession. Noll reported in 1955 that only 14 percent of the colleges that trained teachers required a measurement course and that only 10 percent of the states specified that such a course had to be taken to gain teacher certification. Citing these numbers, Mayo [1959] called for an increase in such courses as requirements. Durost [1959] further darkened the waters by reporting that teacher training institutions were not providing adequate training in this area, largely because courses in testing were being offered "by persons with little background in public education and little understanding of the classroom needs and problems in this area..." [p. 31]. Furthermore, Durost found that teachers were afraid of testing courses and were therefore avoiding them. Teachers' fear of such courses had apparently not abated by 1967, for Mayo reported in that year that courses in statistics were still fearsome prospects among teacher candidates.

In 1972, Roeder reported on a survey of 916 presidents of teacher training institutions. In that survey it was found that 57.7 percent of the institutions required no coursework in evaluation, while 12.1 percent required a one- or two-semester hour course, 17.8 percent required a three-semester hour course, and 1.4 percent required four or more semester hours in evaluation. In addition, 7.2 percent of the institutions required a course in which evaluation was a major component. This state of affairs appears to be an

improvement over the 1955 figures reported by Noll, but Roeder still concluded that teachers were not qualified to use tests.

The situation in the teacher training institutions was illuminated somewhat by Gullickson in 1986. He found a significant misalignment between what college professors emphasized in their measurement courses and the topics for which teachers perceived a need. According to teachers, professors overemphasized statistics and underemphasized non-test methods of evaluation (e.g., observations) and the use of assessment results for instructional planning. Interestingly, Airasian [1991] supported teachers' perceptions of preservice needs by calling for an increased emphasis on informal assessment and a decreased emphasis on standardized achievement tests. Similarly, Stiggins [1985] has called for an increased emphasis among the measurement community on perceived teachers' needs, especially in the area of classroom testing.

As for state requirements regarding testing courses for teacher certification, a recent study [O'Sullivan & Chalnick, 1991] found that only 15 states required such coursework for certification. The conclusion may be inescapable that repeated calls by psychometricians for an increased emphasis on testing before the teacher enters the classroom are largely unheeded and may be in vain.

It may be more sensible to advocate that testing issues be more fully covered as inservice offerings. This has been the message of Traxler [1960], Hagen and Lindberg [1963], and Fleming [1971]. For example, Hagen and Lindberg called it "unrealistic" to assume that teachers could learn all they would need to know about testing during their preservice training; less practically, they also called for the presence of a test coordinator in each school system.

Another suggestion has occasionally been made. As long ago as 1941, Jones and Galbraith concluded that teachers had an impossible task. They had to "interpret [test] results as precisely as the psychologist without the advantage of the psychologist's training and without recourse to his immediate advice" [p. 225]. Since teachers did not, in the

authors' opinion, have the means to acquire the psychologist's learning readily, they suggested that test manuals must become the place where teachers could receive all the information they would need to interpret the tests, presented in language that they could understand. As they put it: "...more extensive and wiser use of standardized tests awaits the development of test manuals in which the emphasis is upon interpretation" [p. 227].

More recently, Rudman [1987] focused on test reports as an important link between classroom instruction and tests at the secondary school level. He examined recent changes in the content and format of score reports provided for the supposed use of teachers by test publishers, such as the "Achievement Ability Comparison" provided with the Stanford Achievement Test, which compares the achievement of a student as measured by the Stanford with his or her "ability," as measured by the Otis-Lennon School Ability Test, and the analyses of student performance on clusters of content (e.g., homophones) provided by several publishers in skill-referenced sections of score reports. He concluded that score reports had clearly improved over the years, but lamented the fact that teachers and administrators seemed to be unfamiliar with these newer reports and that, at least at the secondary school level, score reports were mostly read by counselors and ignored by teachers.

Given that teachers and other educators generally appear to lack what many psychometricians regard as sophistication, or even adequate competence, in testing issues and concepts, and that repeated pleadings for increased emphasis on measurement in preservice and inservice teacher education have not been entirely fruitful, a more intense focus on the score report, the primary mode of communication between test and teacher (and ultimately, teacher and other audiences), may be advisable. In this vein, Tittle [1989] called for more intensive and extensive involvement of teachers not merely as recipients of psychometric knowledge and concepts in their teacher preparation institutions, but as codesigners (with psychometricians) of assessments, including the nature and formatting of

score report information, to bridge the gap between teachers' and psychometricians' understandings of the meanings that can and should be derived from test results.

The Reporting of Test Information

Surprisingly little research attention has been paid to the score report in all the years of standardized achievement testing. Relatively standard approaches to the reporting of test information have evolved over the years, and the report contents and formats of the major test publishers are generally similar, but a research base for the currently prevalent approaches seems not to be available. Ideally, the purposes for testing that are recognized as valid and useful by teachers should drive the nature of the score reports that are provided to teachers. With firmly established and accepted purposes in hand, the content of the score reports for each purpose should be a matter of teacher need, opinion, and preference, mitigated by test content and psychometric standards and limitations. Next, the wording of the reports for each purpose should be determined to be "teacher-friendly," ideally by reference to teacher preferences and opinions. And finally, the presentation and format of each report should, while maintaining accuracy, be helpful, visually appealing, and clear to teachers, and should meet commonly accepted standards of graphic communication [e.g., Rogers, 1961; Schmid & Schmid, 1979; Tufte, 1983, 1990]. Although some of these issues have been addressed by individual researchers in the field, no systematic body of studies appears to have been reported in the literature.

The Content of Reports

Some discussion in the literature has been devoted to the choice of score type (i.e., percentile rank, grade equivalent score, stanine, etc.) most suitable for reporting test results for norm-referenced test interpretations. Not only have the merits and demerits of the

various score types been discussed, but the advisability of using score points vs. score bands has also been a topic for consideration.

In an early study of the effect of using percentile bands instead of ranks as a means of emphasizing that test scores contain error and uncertainty, Morse [1964] presented psychology students with scores expressed as percentile ranks, narrow percentile bands (i.e., ± 0.5 sd), and wide percentile bands (± 1 sd). He found that the students tended to rate bands as closer to the mean than ranks, and narrow bands as closer to the mean than wide bands. On the basis of these findings, Morse advised against bands, stating that they tended to make the person receiving the score think of himself or herself as average rather than above average or below average.

Nevertheless, the weight of the literature is in favor of the use of bands to preclude overinterpretation of test scores. Goldman [1972], Lyman [1974], Mehrens & Lehmann [1985, 1987], Huba [1986], and Hanna [1988] are typical of the widespread support for percentile bands in the literature. Lyman [1974] and Mehrens & Lehmann [1985] in particular stressed that percentile bands were the preferred way to communicate precision information to nonmeasurement specialists; and since teachers are clearly in this category, it may be expected that percentile bands are most appropriate for them.

The issue of precision is the primary reason for interest in stanines, which are nine-unit normalized standard scores reflecting a normal probability distribution [Lyman, 1974]. Because so many raw scores are grouped together into one stanine unit, it is difficult to overinterpret the precision of test information, and significant score differences are likely to be reflected in different stanine units. Moughamian [1965] and Rost [1973] recommended stanines as part of a trend toward coarser units of measurement, and Noeth [1976] favored their use in counseling situations. However, Mehrens and Lehmann [1985] considered stanines less desirable than percentile bands, largely because their meaning was not as readily apparent to the lay person as the concept of percentiles.

One type of score that is widely used but that has encountered considerable disapprobation in the literature is the grade equivalent score. Hostrop [1966] called for the abolition of these scores, in favor of percentile ranks, because of the unstable meaning of grade levels across a nation of different schooling policies and definitions of grades. Goldman [1972] disapproved of their use in counseling, and Reynolds [1981] called the standard, federal definition of learning disability ("two years below grade level in a particular subject but not in others") a fallacy. Reynolds based his objection on the unevenness of the grade level metric from grade to grade, which results in irregularity and distortion in the magnitude of aptitude/achievement discrepancies required for diagnosis of a learning disability across grade levels.

Campbell [1981] objected to grade equivalent scores primarily because they are difficult to explain to parents, who tend to be hurt if their child is below grade level (a necessary condition for about half of the students in the average classroom) and eager to advance their child if he or she is above grade level (a necessary condition for the other half of the students in the average classroom). Because of the confusion inherent in the concept of grade level, Campbell preferred percentile ranks. Mehrens and Lehmann [1985] cited a number of disadvantages of grade equivalent scores in their description of various score types, ultimately preferring percentiles. Green [1987] was more neutral toward grade equivalents in his review of score types, but Genck [1989] was unambiguously opposed to them as indicators of school progress.

The most significant proponent of grade equivalent scores is Hoover [1984], who favors them for measuring educational development in the elementary schools.

Many other types of scores are available for norm-referenced interpretations, including mental age scores, normal curve equivalents, and standardized z and T scores, and each has its uses [Mehrens & Lehmann, 1985, 1987]. In fact, many of these score types appear on the score reports of the major achievement test batteries, probably because each has its advocates and publishers attempt to meet as wide a range of needs as possible.

But the most commonly used score type for reporting individual scores and group (e.g., class) scores is probably the percentile rank, with a confidence band around it to indicate the standard error of measurement. Such a band is an effort to meet professionally accepted standards for admitting score imprecision, as stated in the *Standards for Educational and Psychological Testing* [AERA et al., 1985]. Computation of one or more bands (reflecting differing standard errors of measurement at different score points) is not trivial, but adequate guidance is available in the literature [e.g., Hanna, 1988; Kolen, 1988; Schulte & Borich, 1988]. Publishers' technical manuals ought to include a description of the methods used to compute and report confidence bands, since several methods are in use.

Ever since instructional objectives came into use, it has been recognized that a desirable characteristic of test score reports would be the reporting of scores according to objectives. Even before criterion-referenced testing became common, those writing in the area of test score use suggested ways for the teacher or the school to define more precisely the relatively broad content of standardized achievement tests and to interpret students' scores at the skill or objective level [e.g., Cox & Sterrett, 1970; Ladd, 1971; Walker, 1968]. Later, calls for truly diagnostic testing became common [e.g., Anttonen & Fleming, 1976; Rost, 1973], and when publishers responded by listing the skills underlying their achievement test batteries and keying them to test items, prescriptions for the use of such "skills profile sheets" began to appear [e.g., Boegli, Whately, & Ward, 1977].

Another development that facilitated the advocacy of diagnostic score reporting was the more general availability and use of the computer in educational settings. As early as 1970, Roberge and Kubinieć developed a FORTRAN program to report student scores at the objective level on teacher-constructed multiple-choice tests; a significant aspect of this program was the reporting of results in narrative format. A similar program, called "Diagnose," designed to report criterion-referenced test results, was described by Furlong and Miller [1978]; this program also used nontechnical prose to report the student's performance relative to others taking the test, the questions answered incorrectly and the

correct responses to them, objectives in which the student's performance was deficient, a list of materials for further study, and optional messages to be supplied by the teacher. Summaries were also provided to the teacher.

Killian [1983] foresaw that microcomputers, then emerging, would change testing, especially by permitting districts to do their own scoring and create their own score profiles. Lenke and Beck [1980] predicted that computers, together with criterion-referenced testing technology, would permit the generation of more instructionally useful score reports, including narrative formats, objective-based reports, and scores that revealed what the student could do rather than merely his or her relative position in the class or in terms of a norm group.

The topic of narrative score reports bears special note. Clearly considered highly desirable by many writers, such reports were the subject of a study by Mathews [1972,1973]. Mathews had teachers analyze the Iowa Test of Basic Skills into 25 skill groups and devise a series of narrative descriptors for varying performance on the items associated with the groups (e.g., excellent, quite strong, rather weak). The teachers also developed a set of practical suggestions for teachers to use in instruction (e.g., suggestions for increasing vocabulary). Reports on individual students and the class were prepared for teachers and, covering their children only, for parents. Mathews then presented the ITBS results to 52 teachers in 16 schools, using both the traditional format and the narrative format. The narrative format was greatly preferred by the teachers as being more accurate, meaningful, and sufficient for most uses; it significantly outperformed the traditional format in 15 of 18 comparisons. To this evident enthusiasm for narrative score reports Roid [1984] added a note of caution, reminding those who write the scripts for such reports of the necessity for using empirically validated decision rules and descriptors rather than private and subjective narratives. Other cautions were mentioned by Mehrens and Lehmann [1987].

Suggestions for the reporting of scores according to objectives and in narrative format have been heeded by test publishers. It is now commonplace among the major achievement test batteries to provide a list of skills that underlie test content and a summary of student and class performance relative to those skills. It is also common, at least as an option, to provide narrative-format score reports, usually for parents, but for teachers who want them as well.

In addition to score type, skill area, and narrative-format scoring, other suggestions for changing the content of achievement test score reports have been made over the years. Betts [1950] expressed the desire to have scores reported in such a way that aptitude (in this case, IQ) and achievement could be compared for individual students. Similarly, Hall [1954] suggested that each student be scored on an "Index of Studiousness," expressing the ratio between the student's educational age and mental age. However, Ebel and Hill [1959] expressed considerable doubt about methods for identifying "overachievers" and "underachievers," such as these suggestions would do, stating that most discrepancies between aptitude and achievement could probably be dismissed as measurement error. However, this idea has not been entirely discredited, and the possibility that the Stanford Achievement Test affords of comparing students' aptitude and achievement because of its common norming with the Otis-Lennon School Aptitude Test has been cited as an advantage of the Stanford by Mehrens and Lehmann [1987] and Rudman [1987].

The reporting of change or gain scores for students is another content-related suggestion that has occasionally arisen, despite the difficulties and ambiguities of using such scores. Rapp and Haggart [1973] proposed a graphic way to estimate expected gain scores for each student, using grade equivalent scores divided by 8 to approximate the average rate of gain expected over the eight years from kindergarten to grade seven. Maxwell and Howard [1981] also proposed rethinking change scores, arguing that they could be helpful if carefully interpreted, and not merely misleading.

An interesting proposal related to individual gain scores was made by Singer and Dreher [1983]. They noticed that parents tended to be confused rather than comforted if their children maintained their position in terms of percentile ranking from year to year, interpreting a 39th percentile one year and a 37th the next as a lack of progress rather than the fairly steady progress it actually represents. Singer and Dreher proposed that a self-comparison for each student, designed to show individual growth over the previous year, be presented along with results from the current year's testing. They suggested that this be done by having schools administer the former year's test (e.g., the third-grade test) a second time to students in the following year (when the students would be in the fourth grade), and presenting three scores: each student's past (i.e., last year's) performance on the third-grade test, his or her *current* performance on the third-grade test (both of these expressed as a percentile rank in comparison with the third-grade norming sample), and his or her current performance on the fourth-grade test (expressed as a percentile rank in comparison with the fourth-grade norming sample). Thus Harry's fourth-grade 37th percentile might be arrayed against both his (last year's) third-grade 39th percentile and his (this year's) 60th percentile achieved on the third-grade test in comparison with third graders in the norm group. This content and format change, together with the tactic of showing the hardest item that the student got right in each year, was overwhelmingly preferred by a mix of 42 teachers and administrators in the study. The Singer & Dreher proposal was endorsed by Flood and Lapp [1989] in an article on the desirability of reporting reading progress to parents; Flood and Lapp also suggested showing students' progress from one level of reading material to another as a visual way to signal progress to parents.

Cummings [1981] and Cummings and Stinard [1983] advocated "student-centered test interpretation" as an active technique for explaining test scores to students. In both articles the authors found the ITBS "Pupil Item Response Record," which reveals the student's response choice for every item, to be an effective device for implementing this technique.

Curtis and Glaser [1983] complained that reading achievement tests did not adequately accord with current reading theory and that the heterogeneity of the items on standardized reading tests "makes it unclear what scores on those tests mean" [p. 143]. More a criticism of test content than of reporting format, their article recommended improvements in the way decoding, semantic knowledge, comprehension, and discourse analysis were measured. In a similar vein, Roeber and Dutcher [1989] discussed Michigan's approach to reporting reading scores on that state's innovative assessment of students. Score reporting at the individual level includes scores on constructing meaning, knowledge about reading, attitudes and self-perceptions about reading, and topic familiarity, and, within "constructing meaning," on the ability to construct meaning by making text-based, intersentence, and beyond-text inferences. It is unclear how helpful Michigan teachers have found this theory-based method of score reporting.

Other suggestions for changing the content of score reports will be briefly summarized here, since they would entail substantial revisions in test development methods on the part of test publishers. Hambleton [1980] proposed a way for item response theory to help test publishers report an accurate estimate of an examinee's true ability relative to a well-defined domain of content, based on the examinee's score on any reasonably-sized subset of items drawn from that domain, even if (as is often the case in norm-referenced tests) that subset of items was not truly representative of the broader domain. Lenke and Beck [1980] also predicted that item response theory would play a greater role in the development and interpretation of standardized achievement tests.

Harnisch [1983] discussed the limitations of global summary scores in terms of the many ways a person could achieve a raw score of N from a set of items, and the different meanings that each score of N could conceal. He proposed using Student-Problem Curve Theory to interpret, by computer, different item response patterns actually achieved by examinees in terms of the characteristics of groups of students displaying similar patterns. Analysis of error patterns was also discussed by McArthur and Choppin [1984] as a

possible future trend in test construction and interpretation. In a similar vein, Birnebaum and Shaw [1985] proposed using task specification charts, containing a great deal of detail about student misconceptions, to develop items, and then interpreting students' performance in terms of their error patterns. Such theory-based methods of test construction may indeed be a future trend, and if so, they may permit far greater precision in the reporting and interpretation of test scores.

Bruno, Holland, & Ward [1988] and Bruno [1989] recommended the use of "modified confidence-weighted—admissible probability measurement" evaluation procedures, instead of simple right-wrong scoring, to increase the information derivable from tests. Such procedures entail having students select confidence ratings relative to their knowledge of each test item; their ratings are then used to interpret their knowledge of the domain of content covered by the items. Because of the difficulty of understanding the procedure and its impracticability, it is unlikely to be applied in the settings in which standardized achievement tests are used.

On a less exotic note, the desire of school districts and states to combine norm-referenced and local curriculum-referenced testing for score interpretation purposes (i.e., to combine the benefits of comparison against a local curriculum with those of comparison against a national norm group) was reported on by Linn and Hambleton [1991]. Despite many dangers in modifying standardized test batteries to accommodate local curricula and provide locally relevant data, Linn and Hambleton reported that this is likely to be a growing trend and proposed recommendations to make it less hazardous.

Language and Format of Score Reports

Few recommendations relating to the specific language to be used to communicate information to teachers are presented in the literature, beyond the usual admonition to use nontechnical terms and everyday language as much as possible. This is certainly sensible

advice, but no empirical studies were found relating to the comparative comprehensibility to teachers of various verbal descriptions used on score reports.

In Great Britain, where major changes in examination practices are occurring with great rapidity, similar admonitions regarding the language of score reports can be heard. A (1991) British Department of Education pamphlet recently focused on the annual reports to parents that would soon issue from a newly expanded national testing program. These reports were expected to be free of jargon, succinct, and "written with the reader in mind" so as to avoid overloading parents' capacities for understanding. In addition, they were expected to highlight positive achievement while identifying weaknesses and suggesting action for future improvement. A more unusual suggestion was that the reports should leave space for parents to write down their own comments in preparation for meetings with teachers [Marston, 1991]. Perhaps these suggestions could be usefully applied to the reports destined for teachers as well as parents.

One other issue related to the language of score reports—the use of narrative score reports—which has received favorable ratings from teachers in at least one study [Mathews, 1973], was discussed above.

Graphical presentation of score information has also received relatively scant attention in the literature. Only two graphical formats are generally discussed: profiles and bands. For example, Putt and Ray [1965] recommended using a student profile showing, for a set of subtests, student expectancy levels (e.g., a grade level), the class average, and the individual scores. Gardner [1977] elaborated on these ideas by presenting options for visual arrays of profile data (including bars, lines, and truncated bars showing discrepancy from the median) and by offering cautions about exaggerations that are a danger inherent in graphics. Hoover and Fleetwood [1977] extended the profile idea to apply to central office uses by arraying school buildings instead of subtests along one of the axes. National norms and local averages are also arrayed on the chart. Bohning [1979a, 1979b] suggested a way for teachers to create their own consolidated profiles of students, and Anastasi [1985]

cautiously endorsed profiles while citing some of the reliability problems associated with subscores.

As for confidence bands, these may be represented by solid lines, rows of x's, rows of stars, solid or empty boxes, hourglass-shaped figures, box-and-whisker plots, or boxes with subtest or skill names inside them. Examples are available in Cunningham [1968], LeSage [1973], Lapointe [1987], Tukey [1977], and Swain [1982].

Prescott [1971] suggested a table array for criterion-referenced tests, in which students are arranged on the vertical axis and skills on the horizontal. Then x's are used to mark skills not mastered by individual students. By reading the array horizontally, the teacher can assess student performance; by reading it vertically, the teacher can identify skills that appear to need more instruction and the students who need it. Such an array is now common in the skills portions of the score reports of the major achievement test batteries.

A more graphically-based variant of the above table array is in use on the CTBS/4 "Objectives Performance Report." Student names appear along the top of the report in narrow columns; objectives to which test items are keyed appear in rows on the left side of the report. The intersections of rows and columns are filled with blank circles, half-filled circles, and filled circles to indicate, respectively, non-mastery, partial mastery, and mastery of objective content. Individual student profiles can be read from the report by trailing the eye downward from the student's name, while a profile of class performance on each objective can be read by scanning horizontally along the rows of the form. The idea behind the format is to facilitate the identification of individual student needs, the formation of temporary work groups of students, and the identification of skills in which students need more or less instruction.

Beyond these simple ideas for graphical presentation, the literature on educational testing is largely mute about graphics. However, methods of arraying data, especially multivariate data, have been explored in other bodies of literature, and it is possible that

some ideas that seem odd because they are foreign to educational testing could be transplanted from other areas. For example, Prediger [1971a, 1971b] and Sprinthall [1967] have suggested using discriminant analysis to locate typical score profiles in a Cartesian plane, and then to array individuals on the plane as a visual way of indicating the profiles(s) they are closest to matching; while this technique was suggested for guidance counseling and used with interest inventories, it is possible it might have an application in educational testing.

Wainer and Thissen [1981] offered an exotic set of graphic techniques, including star diagrams in which stars inscribed in circles could be used to indicate relative quantities of several variables, each represented by one of the arms of the star. The resulting shapes give a sort of polar profile of the mix of variables possessed by each individual. Perhaps such a technique could be used for subtest scores on achievement test batteries. An even more unusual technique reported in Wang [1978] involves "Chernoff faces," cartoon faces in which the size of various features (e.g., nose, mouth) indicates the quantity of several variables; these faces are reported to be useful for quickly locating "family resemblances" between individuals, as well as individuals who are clearly outliers. Perhaps they could find a use in reporting subtest scores.

While these ideas may not be practical, it may be fruitful to explore their potential use with teachers. Certainly it can be said of current score reports from the major standardized achievement test batteries that they present a vast amount of information in dense, largely nongraphical formats. It may be the case that the density of information is useful; however, it may also be the case that it is overwhelming and confusing to the teacher. Perhaps greater use of graphical presentations would be helpful to the teacher in the difficult job of test score interpretation and use.

Conclusion

It should be apparent from the above review of the literature that teachers are essential participants in score interpretation for standardized test batteries, may be underprepared for their technical role, may be helped by score reports that are designed to accommodate their needs and preferences relative to the purposes for which such tests are useful to them, and may be helped by further attention to their preferences regarding the content and format of score reports. It should also be apparent that work is needed on establishing what teachers' preferences are in these matters.

CHAPTER 3 RESEARCH METHODOLOGY

Introduction

The review of the literature presented in Chapter 2 underscored both the central role of the elementary school teacher in using the results of standardized achievement test survey batteries and the relative lack of research attention paid to the needs, perceptions, and preferences of teachers regarding the content and format of the score reports that permit the use and interpretation of SATB results. It was the purpose of this study to contribute to reducing this research deficiency by addressing to elementary school teachers questions designed to elicit their opinions about standardized achievement test survey battery score reports: their purposes, content, formats, and interpretations.

Overview of the Study Design

This study gathered the opinions of a sample of public elementary school teachers in grades 1 through 8 regarding the purposes of SATB score reports in general and the purposes of particular examples of SATB-type score reports. In addition, the study examined teachers' opinions about the appropriateness of classroomwide and individual academic interpretations of score reports of the sort that are typically made by teachers and others. The method used to gather the information for this study was a questionnaire.

Content of the Questionnaire

The questions on the questionnaire relating to the general purposes of score reports were drawn from the literature on score reporting cited in Chapter 2. For the questions that would pertain to particular score report examples, two commonly used types of score reports were used as models: the class-level score report, which reports test information about an entire classroom of students, and the individual student score report, which

focuses on one student's performance on a test as compared with other students in a larger sample. Both types of score reports are widely used and interpreted by classroom teachers.

Because the study concerned the format and presentation of score reports, as well as their informational content, each of the two general types of score reports was presented in two formats. The class-level score report was presented in a *numerical format*, in which all students' scores on the test and on the skills and subskills covered by the test are reported numerically, and a *graphical format*, in which students' scores are presented largely in graphical format. For this study, the numerical format used was similar to reports produced in conjunction with the ITBS and the graphical format used was similar to the CTBS format described above, involving open, half-filled, and fully filled circles to represent different conditions of mastery of test content. A sample of these two formats of a class-level score report may be found in Appendix A, which contains a complete Form A questionnaire.

The individual student score report was presented in a *narrative format*, in which test information is described in paragraphs of text beneath an overall summary (which includes several kinds of scores, such as grade equivalents and national percentiles, together with confidence bands) of the student's performance on broad areas of the test, and a *numerical/pictorial format*, in which test information is presented numerically and through skill- and subskill-level confidence bands beneath an overall summary identical to the one on the narrative report. For this study, the narrative format was similar to reports in use with the SAT, the CTBS, and the CAT, and the numerical/pictorial format was similar to CAT and CTBS reports. A sample of these two formats of an individual student score report may be found in Appendix B, which contains a complete Form S questionnaire.

In addition to the opinions of teachers regarding score reports, personal background information was also sought regarding respondents' experiential and demographic characteristics that might be helpful to the researcher in understanding the nature of the study sample and that might shed light on one question of particular research interest in this

study: the effect, if any, of training in testing and measurement on teachers' opinions and, particularly, on their interpretations of score report information. To this end, three of the background questions asked respondents how many preservice courses, inservice courses, and workshops they had taken that addressed testing and measurement issues as either the sole focus or a major focus.

Sample

The sample of teachers used for the study was intended to be diverse and broadly reflective of public school teachers in the United States in grades kindergarten through eight. To this end, three states diverse in terms of geographical region, population size and composition, and educational history and structure were targeted for the study: Texas, Illinois, and Massachusetts. Elementary school teachers from a variety of schools in these states were invited to participate in the study; the sample of actual respondents was voluntary. Details regarding the method used to contact potential respondents and the nature of the actual respondent sample will be presented later in this chapter and in the next chapter. In all, 671 usable responses were received and are reported in this study: 231 from Texas, 297 from Illinois, and 143 from Massachusetts.

Data Analysis

A variety of analytic strategies was employed to explore and interpret the data for this study. Descriptive statistics were computed to characterize the sample, including frequency distributions for every background question. Means and standard deviations were calculated for responses to the general opinion questions, the score report-specific opinion questions, and the score report-specific interpretation questions. Patterns of responses to the general opinion questions were examined for unusually strong opinions either in favor of or in opposition to proposed purposes for score reports. For the score report-specific opinion questions, response patterns were also examined, as were comparisons of

respondents' opinions regarding the purposes of one format of score report vs. the other. For this analysis, a repeated measures *t* test was applied to the differences in means for each set of two corresponding purpose questions; significant ($p < 0.05$) differences are reported and discussed in Chapter 4.

For the interpretive questions, the degree of respondents' agreement with each interpretation is reported as a mean ranging potentially from 1.00 (strong disagreement) to 5.00 (strong agreement), and the overall degree of respondents' agreement with these interpretations is compared between score report formats (i.e., numerical vs. graphical; narrative vs. numerical/pictorial). In addition, since each interpretation was intended to be to some extent an overinterpretation from a psychometric perspective, respondents' ratings are compared with a small sample of psychometricians' ratings on these questions in terms of degree of agreement with each interpretation.

For all opinion and interpretive questions, background variables that might have contributed significantly to the variance in responses were explored through general linear model (GLM) analysis of variance (ANOVA), using the SAS statistical package. Statistically significant ($p < 0.05$) *F* values were examined and follow-up comparison tests, using the Tukey studentized range test procedure for multiple comparisons, were applied to each comparison.

Since responses to the interpretive questions were of particular research interest in this study, the background variables that related to training in testing and measurement issues were the focus of further study. Correlations between these three background variables and the interpretive questions were calculated and examined in several ways, using several combinations of the raw training data.

Each of these aspects of the study—the questionnaire, the sample, and the data analysis—is described in greater detail in the following sections of this chapter.

The Survey Instrument

The questionnaire developed for this study was intended to gather three types of information from teacher respondents. The first pertains to the general purposes for which SATB score reports are viewed as providing useful information. The second pertains to the perceived usefulness of particular kinds of score reports (i.e., class-level reports and individual student reports) to the classroom teacher; for this section of the questionnaire two different formats for each type of score report were presented to respondents. The third type of information pertains to the interpretations that teachers make regarding the information on particular kinds of score reports; again, respondents were asked to consider two different formats for each type of score report.

In addition to these three types of substantive information, the questionnaire was designed to gather background information from the respondents relating to their teaching setting, their teaching experience, their training in testing and measurement issues, their gender and ethnicity, and other characteristics.

Questionnaire Design

Initial design of the questionnaire involved seeking informal input from teachers and former teachers by means of face-to-face interviews and telephone conversations. In addition, input from faculty on the researcher's doctoral committee was sought. Both sources of input were applied to the goals for the instrument, which had been set based on the review of the literature summarized in Chapter 2.

An initial draft of the instrument produced a questionnaire that contained a first section of background questions, a second section of questions relating to the purposes of SATB score reports in general, and a final section in which four different score reports were presented (two formats of a class-level score report and two formats of an individual student score report), with questions relating to each one. This version was 13 pages long, plus a cover page.

Reactions to the initial draft produced several ideas that were incorporated into a second version, including moving the background questions to the end of the questionnaire and, most important, shortening the instrument, which was taking volunteers nearly an hour to complete. The strategy for shortening the questionnaire was to divide it into two separate questionnaires, each presenting (in addition to the background and general purpose questions) two formats of one basic type of score report (i.e., one class-level score report in two formats in one questionnaire and one individual student score report in two formats in the other). The result was an instrument of nine pages (plus a cover page) in three sections: 20 general purpose questions, 30 score report-specific questions (15 for each of two formats of score report—nine relating to purposes and six relating to interpretations), and 13 background questions. In addition, each questionnaire contained an implicit background datum—state of teaching assignment (TX, IL, or MA)—indicated by the color of the paper on which the questionnaire was printed.

Because each form of the questionnaire contained two score reports in different formats, the possibility that the order in which the formats were presented might affect respondents' opinions was considered. As a result, each basic form of the questionnaire was divided into two forms, one with one format presented first and the other with the second format presented first, so that any potential order effect could be assessed. This division resulted in four forms of the questionnaire: Form A and Form B (both containing class-level score reports in counterbalanced order) and Form N and Form S (containing individual student score reports in counterbalanced order).

The content of the four forms is summarized below:

- Form A: class-level, with numerical format first;
- Form B: class-level, with graphical format first;
- Form N: individual student, with narrative format first;
- Form S: individual student, with numerical/pictorial format first.

Questionnaire Content

The questionnaires contained three sections, the second of which differed from form to form, but the first and third of which were identical. Each section is described below.

Section I

The first section of all forms of the questionnaire presented the same 20 statements, each one to be considered in terms of a five-point Likert-type scale that included the categories Strongly Disagree (SD), Disagree (D), Neutral (N), Agree (A), and Strongly Agree (SA). In addition, a No Opinion (n/o) option was presented for each statement. The 20 statements are purposes commonly cited for SATB score reports in the research literature or in SATB test publishers' written materials relating to their products (see Chapter 2).

The 20 statements in Section I are these:

(SATB score reports provide useful information for...)

1. helping schools make decisions about placement of individual students into permanent instructional groups (e.g., homogeneous ability groups).
2. helping teachers make decisions about placement of individual students into *temporary* instructional groups (e.g., cooperative learning groups, groups for enrichment or remedial work).
3. helping teachers diagnose individual students' strengths, weaknesses, and needs.
4. helping teachers keep the pace and level of instruction "on track" with national expectations.
5. enabling teachers to measure individual students' growth in particular skills.
6. enabling teachers to measure individual students' growth in overall subject areas (e.g., math, language arts).
7. enabling teachers to measure group achievement in particular skills over time.
8. enabling teachers to measure group achievement in overall subject areas over time.
9. enabling teachers to plan instruction that is tailored or adapted to individual students' needs.

10. helping teachers establish individual students' grades in class.
11. enabling schools to make promotion/retention decisions for individual students.
12. helping teachers and schools evaluate the effectiveness of the curriculum or of curriculum materials.
13. helping teachers evaluate the effectiveness of their own instructional approaches and strategies.
14. helping students gain personal insight into their strengths and weaknesses.
15. helping teachers gain unexpected insights into particular students' hidden talents, achievements, or interests.
16. enabling teachers to compare individual students' aptitude and achievement levels.
17. helping teachers eliminate potential sources of personal bias in evaluating their students' abilities by providing an objective form of information on student achievement.
18. helping teachers explain individual student achievements and needs to parents.
19. enabling administrators to compare varying programs or approaches being implemented in different classrooms or schools.
20. helping administrators evaluate the performance of individual teachers.

Section III

Section III, the other section of the questionnaire that was uniform across the four forms, contained 13 background questions in multiple-choice format. The primary purpose of these questions was to describe and characterize the respondent sample. Although the sample is strictly neither a random sample nor a stratified random sample, it was intended to be a diverse sample with characteristics similar to those of larger teacher populations.

The background variables included were grade level(s) taught, setting of primary assignment, population of the municipality of primary assignment, years of teaching experience, gender, and ethnicity. An implicit background variable, tracked through the color of the paper on which the questionnaire was printed and the method of distribution and collection, was state of teaching assignment (i.e., Texas, Illinois, or Massachusetts).

In addition to these variables of interest, information pertaining more directly to the study was collected through the background questions. Three of the questions related to the amount of training in testing and measurement issues possessed by the respondents; these questions were included to address the hypothesis that teachers with different amounts of psychometric training would answer opinion and interpretive questions about SATB score reports in systematically different ways.

Finally, several background questions addressed, for descriptive purposes, respondents' opinions about the usefulness of their testing and measurement preparation in the performance of their jobs and the frequency of their drawing on testing and measurement knowledge in their jobs. Two final questions asked respondents which SATBs they were familiar with and which ones their schools used currently.

The Section III questions are reproduced here:

Section III: Background Questions

(Please circle one response letter unless otherwise specified.)

- To which grade levels are you currently assigned as a teacher? (Circle response letters for all levels that apply.)

A. Kindergarten	D. Grade 3	G. Grade 6	J. Other
B. Grade 1	E. Grade 4	H. Grade 7	
C. Grade 2	F. Grade 5	I. Grade 8	
- Which of the following describes your primary work environment?
 - Self-contained classroom (teaching the same group of students more than one subject)
 - Departmentalized (teaching the same subject to different groups of students)
 - Multi-setting/itinerant (teaching at more than one school)
 - Administrative (e.g., instructional coordinator)
 - Other
- What is the approximate population of the municipality (i.e., town or city) in which your school of primary assignment is located?

A. under 25,000	D. 250,000 to 499,999
B. 25,000 to 99,999	E. over 500,000
C. 100,000 to 249,999	

4. How many years of teaching experience do you have, counting this year?
 - A. fewer than 3
 - B. 3 to 9
 - C. 10 to 20
 - D. more than 20

5. During your *preservice* teacher training, how many courses did you take that addressed testing and measurement issues as either the *sole focus or a major focus*?
 - A. none
 - B. 1
 - C. 2
 - D. 3
 - E. more than 3

6. Since you started to teach, how many *in-service courses* (i.e., for credit, not professional development workshops that carried no credits) have you taken that have addressed testing and measurement issues as either the *sole focus or a major focus*?
 - A. none
 - B. 1
 - C. 2
 - D. 3
 - E. more than 3

7. Since you started to teach, how many *in-service workshops* (i.e., not for credit) have you taken that have addressed testing and measurement issues as either the *sole focus or a major focus*?
 - A. none
 - B. 1
 - C. 2
 - D. 3
 - E. more than 3

8. In general, how useful do you think your preservice and inservice preparation has been for dealing with testing and measurement issues that arise on your job?
 - A. not at all useful
 - B. rarely useful
 - C. sometimes useful
 - D. generally useful
 - E. very useful

9. Approximately how often do you have to draw upon your knowledge of testing and measurement issues as part of your job?
 - A. never
 - B. rarely (1 to 5 times a year)
 - C. occasionally (6 to 15 times a year)
 - D. often (16 to 30 times a year)
 - E. very often (more than 30 times a year)

10. Are you Female or Male?
 - A. Female
 - B. Male

11. What is your racial/ethnic background?
- A. American Indian/Alaskan Native
 - B. Asian/Pacific Islander
 - C. Black, Non-Hispanic
 - D. Hispanic
 - E. White, Non-Hispanic
12. With which of the following standardized achievement test batteries (SATBs) are you familiar? (Circle response letters for all that apply.)
- A. California Achievement Test (CAT)
 - B. Comprehensive Test of Basic Skills (CTBS)
 - C. Iowa Test of Basic Skills (ITBS)
 - D. Metropolitan Achievement Test (MAT)
 - E. Stanford Achievement Test (SAT)
13. Which of the following standardized achievement test batteries (SATBs) does your school currently use? (Circle response letters for all that apply.)
- A. California Achievement Test (CAT)
 - B. Comprehensive Test of Basic Skills (CTBS)
 - C. Iowa Test of Basic Skills (ITBS)
 - D. Metropolitan Achievement Test (MAT)
 - E. Stanford Achievement Test (SAT)

Section II—Purpose Questions

Section II of each form of the questionnaire contained four pages relating to two formats of score reports. On facing pages were one format of a score report and 15 statements (with Likert-type scales) that related to the information on that score report; the second format and its 15 statements followed on the next two-page spread. Of the 15 statements, the first nine addressed uses of the score report in question; these are referred to as the purpose questions and they are identical or virtually identical across the two report formats in any one questionnaire.

For the class-level score report type in Form A and Form B of the questionnaire, the purpose questions were as follows for the first of the two formats (Score Report Sample A):

Score Report A would be useful to a classroom teacher who wants to:

1. compare his or her students' achievement levels to those of students nationwide.
2. understand the instructional needs of particular students.

3. form temporary work groups to focus on individual skill development.
4. distinguish skill areas that need emphasis from those that do not.
5. evaluate his or her own teaching effectiveness.
6. target curriculum areas in which resources and/or teaching methods should be reevaluated.
7. provide feedback to parents on the skills of their children.
8. evaluate the effectiveness of the math curriculum.
9. plan instruction that is tailored to individual students' needs.

For the second of the two formats (Score Report Sample B) in this questionnaire, the purpose questions were the same except for statement 8. Since Score Report Sample B presented class-level information on the reading/language arts portion of a hypothetical SATB, while Score Report Sample A did the same for the math portion, statement 8 of Score Report Sample B read "evaluate the effectiveness of the reading/language arts curriculum."

The purpose questions that related to the other type of score report under study, the individual student score report, were identical from one format (i.e., narrative) of the score report (Score Report Sample N) to the other (i.e., numerical/pictorial—Score Report Sample S). The purpose questions for Format S of this type of score report are these:

Score Report S would be useful to a classroom teacher who wants to:

1. compare this student's achievement levels to those of students nationwide.
2. understand this student's academic strengths and weaknesses.
3. create an instructional plan targeted to this student's needs.
4. find out the difficulty level of the reading materials with which this student will be comfortable.
5. learn the grade levels at which this student is performing in the skill areas covered by the test.
6. help this student understand his own test performance.

7. discuss this student's test performance with his parents.
8. evaluate the effectiveness of instructional strategies, curriculum, and/or resources now in use in this classroom.
9. set up groups of students to work together on specific skills.

Section II—Interpretive Questions

For each form of the questionnaire, Section II also contained interpretive questions pertaining to the particular score report samples that were part of that form. Because they were related to the content of different score reports, the interpretive questions differed from form to form. The interpretive questions are quoted and discussed below, with reference to each sample score report. Since each interpretive question was intended to be to some extent an overinterpretation of the information on the score report, the discussion explains why it was so intended, the extent to which it was intended to be an overinterpretation, and the results of an informal administration of this portion of the questionnaire to a local group of eight persons with doctoral degrees and experience in testing and measurement (referred to as the local psychometric group, or LPG).

The concepts involved in the interpretive statements relate to such psychometric issues as reliability, appropriate confidence in numerical scores, and the need to apply caution and restraint in drawing conclusions from test data. The statements represent misinterpretations of several kinds: regarding numerical information as precise rather than an approximation within a range of confidence; making strong inferences and sweeping generalizations from data that do not support them; interpreting grade equivalent scores as making a statement about the level *to which* an examinee is appropriately assigned rather than one about the level *from which* examinees in the norming sample came; trusting numerical scores even in a context of student information that contradicts their messages; and placing students into different categories despite evidence, such as overlapping confidence bands, that they might be in the same category. These concepts are well

discussed in the standard textbooks on testing and measurement, such as Anastasi [1988], Brown [1983], and Mehrens and Lehmann [1987].

As should become clear from the following discussion, absolute right/wrong answers to these interpretive questions were neither intended nor achieved. In most cases, the LPG indicated agreement that the researcher's intention to provide overinterpretations of the score reports had been accomplished. But the extent of the LPG's agreement, indicated by its mean rating on each question, varied from question to question, and on four of the 24 questions the LPG contradicted the researcher's thinking and agreed with the supposedly overinterpreted statement.

Sample Score Report A. This score report (see sample in Appendix A), which is the numerical format of a class-level score report, presents information pertaining to the math portion of a hypothetical SATB for each of 26 students, arranged alphabetically, in a fifth-grade class. The interpretive questions are numbered 10 through 15. Each is really a statement preceded by the phrase, "On the basis of Score Report A, it is justifiable to conclude that:" Each statement is quoted below and then discussed.

A10. compared with students nationwide, this class is below average in "Math Concepts" and above average in "Math Computation."

This statement was intended to be a slight overinterpretation of the test results, since the class scores on each subsection of the "Math Concepts" section of the test, which in all comprised 35 items, are a little bit lower than, but in fact quite close to, the national percent correct (i.e., the percent of students in the norming sample who got these items correct) and the class scores on the subsections of the 39 items in the "Math Computation" section of the test are only slightly higher than the scores achieved on these items by the norming sample. Without further information regarding the characteristics of this sample and, especially, the degree of confidence one can have in the local scores on these items (i.e., the reliability of the subtests), it is a bit of an overstatement to conclude that this class

is below or above average on any subtest. The local psychometric group (LPG) was neutral on this statement, giving it a mean rating of 3.00, the midpoint of the scale.

A11. as a whole, this class knows more about concepts related to "Number Systems/Whole Numbers" than about concepts related to "Decimals and Percent."

This statement was intended to be quite an overstatement despite the apparent superiority in class percent correct for "Number Systems/Whole Numbers" over "Decimals and Percent." A careful look at the individual student scores across the rows of the table would reveal that in fact the students' scores on "Decimals and Percent" *varied less around the mean* than those on "Number Systems/Whole Numbers." But on the items for the "Number Systems/Whole Numbers" section of the test, the whole class pattern revealed that while a few students apparently performed well, many others performed poorly. Therefore to conclude that the class as a whole knows more about concepts related to "Number Systems/Whole Numbers" than about concepts related to "Decimals and Percent" is probably too strong an inference. The LPG tended to endorse the researcher's intention by rating this item 2.38, i.e., on the "Disagree" side of the scale.

A12. of the four areas covered by the test, this class needs the most work in "Math Problem Solving."

This statement was intended to be very plausible, but not entirely justified by the data. While students achieved their lowest numerical scores in this area of the test, the scores are in fact quite close to the norm group's scores. Whether the most work is needed in this area, the "Math Concepts" area, or even the "Mathematical Expression" area could not be determined from this score report, in the opinion of the researcher. The LPG concurred, rating this statement 2.63, on the "Disagree" side of the scale.

A13. this teacher spends too much time on "Math Computation" and not enough time on "Math Problem Solving."

This statement was intended to be a considerable overinterpretation. Test results alone tell nothing about the time spent by a teacher on various curricular areas. Such an

inference seemed to the researcher to be unsupportable. The LPG agreed: their mean rating was 1.57.

A14. Luisa Ali needs remedial work on "Fractions."

This statement was intended to be very probably an overinterpretation because Luisa Ali's other scores on this test (visible from reading down her column) were all consistently fairly high; the "Fractions" score was atypical. This was intended to suggest a testing fluke and to induce some suspicion of the test result. The LPG moderately agreed: their rating was 2.88.

A15. Seth Viola knows as much about concepts relating to "Equations" as Theodora Xavier does.

Again, this statement was intended to be a very probable overinterpretation because Seth Viola's "88" in "Fractions" appears in the context of his generally low scores and could thus be an artifact of the test or the testing situation, while Theodora Xavier's "88" appears in a more homogeneous score environment. The LPG agreed that this statement was an overinterpretation, giving it a mean rating of 2.63.

Sample Score Report B. This score report (see sample in Appendix A) is the graphical version of the previous score report, with the same class of students as its focus. However, this report concerns the class's performance on the reading/language arts portion of the test. Many of the same issues are covered on this sample as were covered on the numerical report (Sample A), such as overinterpretation based on too little information or too strong an inference from the data.

B10. compared with students nationwide, this class is above average in "Language Expression" skills.

This statement was intended to be a slight overinterpretation of the test results, since the class scores on "Language Expression" cannot on the basis of this report be conclusively called "above average" for students nationwide without more information than

is given regarding the norming sample and the reliability of the subtest. The LPG did not agree with the researcher's interpretation, indicating weak agreement with this statement by giving it a rating of 3.14.

B11. as a whole, this class has more knowledge gaps on skills covered under "Reading Comprehension" than on those covered under "Reading Vocabulary."

This statement was intended to be a substantial overstatement. The numerical information on these two areas of the test is ambiguous, but the graphical information is less so: the circles indicating mastery level on "Reading Vocabulary" are more often open, or empty (indicating nonmastery), than the circles for "Reading Comprehension," which displays more "partial mastery" (i.e., half-filled circles) of the content. Thus "Reading Vocabulary" displays more and larger knowledge gaps than "Reading Comprehension." Because assessing the information to which this statement pertains would supposedly be facilitated by the graphical format of the score report, the researcher expected that respondents would disagree more readily with this statement on this form than with corresponding "gaps" or "needs" statements on the numerical version of the form (compare statements A11 and A12, discussed above). The LPG did not bear out this expectation. Although they did disagree with the statement (rating = 2.89), they disagreed more strongly with statements A11 and A12 (2.38 and 2.63, respectively).

B12. as a whole, this class needs about the same amount of work on "Passage Analysis" and "Central Idea."

B13. as a whole, this class knows more about skills covered under "Words in Context" than skills covered under "Stated Information."

As with statement 11, these statements were intended to be overinterpretations. In both cases the numerical information was expected to be overridden by the graphical information. If the graphical circles are used, it becomes clear that more work is probably needed on "Passage Analysis" than on "Central Idea" (i.e., there are more open circles) and

that the class probably knows less about "Words in Context" than it does about "Stated Information." Mean LPG rating for 12 was 2.38; for 13 it was 2.88.

B14. the students in this class spend too much time learning grammar and not enough time actually writing.

As with statement A13, this statement was intended to be a considerable overinterpretation. Test results alone do not reveal the activities on which students are spending time. The LPG agreed: 1.88.

B15. compared to Luisa Ali, George Benne is stronger in "Reading Comprehension" but weaker in "Language Analysis."

This statement was intended to be a mild overinterpretation. In both areas, both students are in the same or adjacent mastery categories. Without reliability information it is not strictly possible to conclude that one is stronger than the other in either of the two areas. The LPG was neutral: 3.00.

Sample Score Report N. This score report (see sample in Appendix B) is the narrative version of the individual student score report. A discussion of the interpretive questions follows.

N10. this student has the math problem solving skills of a third grader.

This statement was intended to represent a common overinterpretation of grade equivalent scores. The grade equivalent score of 3.2 on "Math Problem Solving" merely indicates that this student performed about as well on that portion of this test as the students in the norming sample who were in grade 3.2 performed on the same (fifth grade) items. It says nothing about this student's ability relative to the math problem solving skills of a third grader. LPG rating was 2.63.

N11. compared with the nation's fifth graders, this student is above average on the skills covered under "Language Analysis."

This statement was intended to represent a common overinterpretation of national percentiles. This student's 53rd percentile is really an approximation based on a confidence band of 48 to 68. This student *may be* above average compared with this sample (i.e., above the 50th percentile), but he may also be below that percentile. The narrative information ("This student performed above the national average [the 50th percentile] on all reading subtests..."), which is modeled on an actual narrative score report, together with the summary number "53" under "NP" are in conflict with both the numerical confidence band information under "range" and the pictorial confidence band, which clearly extends below the 50th percentile. Mean LPG rating was 2.50.

N12. on the skills covered under "Total Math," this student performed better than 28 percent of the nation's fifth graders.

As with the statement above, this statement was intended to represent a common overinterpretation of national percentiles. But perhaps because of the narrative report's ". . . better than approximately 28 percent of the nation's fifth graders," the LPG tended to agree rather strongly with this statement: 4.13.

N13. this student knows more about the skills covered under "Language Expression" than those covered under "Language Analysis."

This statement was intended not to be supported strongly by the score report. The numerical scores for "Language Expression" at the top of the report were higher than for "Language Analysis," but the confidence bands clearly overlapped and the band for "Language Analysis" even extended higher than that for "Language Expression." LPG rating was 2.88, indicating weak disagreement with the statement, as expected.

N14. this student knows more about the skills covered under "Add/Subtract Whole Numbers" than those covered under "Use of Math Symbols/Terms."

This statement was intended to be unverifiable. Even with the narrative information that "Add/Subtract Whole Numbers" was the only objective on which the student had achieved partial mastery, to conclude that the student was decisively in the nonmastery category on other skills was intended to be an overinterpretation. LPG rating was indecisive: 3.00.

N15. this student is in the "Not Mastered" category on the skills covered under "Multiply/Divide Whole Numbers."

Since this objective was mentioned in neither the upper portion of the report nor the narrative portion, the respondent would be left to infer an answer. As with the objective above, the only information given is that the student is in the "Partial Mastery" category on "Add/Subtract Whole Numbers." To conclude without detailed score information that this student was decisively in the nonmastery category on an unmentioned objective was intended to be an overinterpretation. The LPG did not agree with the researcher's reasoning; its mean rating was 4.13.

Sample Score Report S. This score report (see sample in Appendix B) is the numerical/pictorial version of the individual student score report. It provides more detailed, objective-level information than the narrative report. A discussion of the interpretive questions follows.

S10. this student has the reading vocabulary of a beginning sixth grader.

As with statement N10, this statement was intended to represent a common overinterpretation of grade equivalent scores. The grade equivalent score of 6.1 on "Reading Vocabulary" merely indicates that this student performed about as well on that portion of this test as the students in the norming sample who were in grade 6.1 performed

on the same (fifth grade) items. It says nothing about this student's ability relative to the reading vocabulary of a sixth grader. LPG rating was 2.63, the same as for statement N10.

S11. on the skills covered under "Total Reading," this student performed better than 58 percent of the nation's fifth graders.

This statement was intended to represent a common overinterpretation of national percentiles. This student's 58th percentile is really an approximation based on a confidence band of 52 to 65, and strictly speaking, it pertains to the norming sample, not to all the fifth graders in the nation. However, the mean LPG rating was 3.14.

S12. compared with the nation's fifth graders, this student is in the lowest quartile on the skills covered under "Math Concepts."

Both numerical information (the confidence band running from the 18th to the 30th percentile) and graphical information (the confidence band clearly crossing the line representing the 25th percentile) were intended to contradict this statement, which was based on the NP score of 21. The LPG rating supported the researcher's analysis: 2.88.

S13. this student knows more about the skills covered under "Sentence-level Mechanics" than those covered under "Writing Conventions."

This statement was intended to address the issue of overlapping confidence bands. Raw scores of 76 and 78 are too close to distinguish in the presence of confidence bands that overlap. The LPG concurred: 2.63.

S14. this student knows more about the skills covered under "Math Computation" than those covered under "Math Expression."

Similar NP scores (37 and 34) and identical confidence bands (30 - 44) were intended to lead to rejection of this statement. For the LPG the intention was fulfilled: 1.88.

S15. this student is in the "Mastered" category on the skills covered under "Use of Nouns, Pronouns."

A confidence band clearly crossing the Partly Mastered/Mastered boundary was juxtaposed to a raw score of 78. The confidence band should prevail, from the psychometric point of view. The LPG concurred: 2.50.

Open-Ended Comments

In addition to the structured input of the questions on the questionnaire, respondents were invited to write comments after Section I and on the back cover of the instrument. These comments are briefly discussed in Chapter 4.

The Respondent Sample

The target population for the study was the population of public school teachers in the elementary grades, defined for this purpose as kindergarten through grade eight, in the United States. Although neither a random sample nor a stratified sample of this population was drawn, a diverse sample of teachers from different regions of the United States was sought, with the intention of reflecting to some extent the varied characteristics of the population. To this end three states were selected for sampling, differing in geographical region, demographic composition, and history and structure of public schooling. Texas was chosen as a large southwestern state with a substantial mix of ethnicities, including a sizable Hispanic population, and a history of centralized public schooling. Illinois was chosen as a large midwestern state with a different mix of ethnicities, including a sizable African American population, and a history of moderately decentralized public schooling. Massachusetts was chosen as a small northeastern state with some ethnic diversity and a history of strongly decentralized public schooling.

The target size of the sample was initially 300 teacher respondents, but when the questionnaire was divided into two versions to reduce its length, the target sample size was increased to 400 – 450 because each version of the questionnaire would be completed by about half the sample. It was estimated that 200 – 225 responses to each survey type would yield stable information provided that diversity with respect to demographic characteristics could be achieved in the sample. In fact, 671 usable responses were received, 307 for one survey type and 364 for the other.

Each of the two versions of the questionnaire (one containing class-level score reports and the other containing individual student score reports) was prepared in two counterbalanced forms in which the two sample score reports were presented in opposite order. Distribution of the four forms (A, B, N, and S) of the questionnaire was to be random. Appendix A contains a sample of Form A and Appendix B contains a sample of Form S.

The sample was voluntary. Beyond the professional curiosity and sense of responsibility of public school teachers, no inducement to respond was offered. Because of this fact, a moderately centralized distribution strategy was employed.

Distribution and Collection Strategy

In each state, a group of "survey liaisons" was sought to serve a central role in survey distribution and collection. The survey liaisons—who would generally be teachers themselves, although in a few cases principals, a testing director, and one superintendent served in that role—would agree to receive packages containing a mutually determined number of surveys, to identify and contact as many K–8 elementary school teachers in their school or school district as possible and encourage their participation, to distribute survey materials to the potential respondents, to receive from the potential respondents their completed questionnaires (in envelopes to ensure privacy), and to return the packets of completed questionnaires to the researcher. The packets of survey materials that the liaisons would receive were to contain all four forms of the questionnaire spiraled for random distribution.

Texas Sample

In Texas, the researcher received help from the Director of the Division of Teacher Assessment of the Texas Education Agency and his staff in identifying potential survey liaisons in a variety of towns and cities in Texas. In selecting potential survey liaisons, the

diversity of the desired respondent sample was a central consideration. A list of 23 potential liaisons was compiled, each of which the researcher attempted to contact by telephone and/or by mail. In the end, 16 educators from 12 different municipalities agreed to serve as survey liaisons. Each one was sent the number of surveys that he or she felt could be distributed to teachers in his or her school or school district, plus about 10 percent to cover lost or damaged surveys.

The list of municipalities in the Texas sample and the numbers of survey packets sent in that state is presented in Table 3.1.

Table 3.1
Survey Packets Sent to Texas

Municipality	Number Sent
Alief	40
Arlington	32
Austin	40
Corpus Christi	70
Dallas	52
De Soto	32
El Paso	24
Houston	166
Lubbock	30
Nacodoches	32
San Antonio	24
Waco	32
Total	574

The mailings of sets of materials to the survey liaisons included a checklist; an acknowledgement of receipt form with a return envelope; a copy of the study abstract; a draft of a cover letter to survey recipients from the liaison, which could be used by the liaison if desired; a draft of a follow-up letter from the liaison to recipients, also for

optional use by the liaison; a sheet of suggested distribution, collection, and return procedures for the surveys; several return Federal Express envelopes addressed to the researcher; and the number of survey packets agreed upon, in manila envelopes and spiraled randomly among the four forms (samples of these materials, excluding the envelopes and questionnaires, are contained in Appendix C).

Each survey packet for the potential respondents in Texas included (in addition to one form of the questionnaire) a cover letter from the researcher explaining the purpose and nature of the study; a copy of a letter encouraging teacher participation in the study from Dr. Nolan Wood of the Texas Education Agency, the Director of the Division of Teacher Assessment who had helped through his staff with the identification of suitable liaisons; and a form for respondents to use in requesting an executive summary of the results of the study (samples of these materials are contained in Appendix D).

Of the 574 survey packets sent to the survey liaisons, 231 usable surveys (40.2 percent) were returned to the researcher.

Illinois Sample

In Illinois, the researcher received help from the Assistant Superintendent of Teacher Education and Certification of the Illinois State Board of Education and her staff in identifying potential survey liaisons in a variety of towns and cities in Illinois. In selecting potential survey liaisons, the diversity of the desired respondent sample was a central consideration. A list of 16 potential liaisons was compiled, each of which the researcher contacted by telephone and/or by mail. In the end, 11 educators from eight different municipalities agreed to serve as survey liaisons in their schools or school districts. Each one was sent the number of surveys that he or she felt could be distributed, plus about 10 percent to cover lost or damaged surveys.

The list of municipalities in the Illinois sample and the numbers of survey packets sent in that state is presented in Table 3.2.

Table 3.2
Survey Packets Sent to Illinois

Municipality	Number Sent
Belleville	98
Cairo	100
Carbondale	100
Chicago	104
East St. Louis	40
Elgin	60
Elk Grove Village	30
Teutopolis	32
Total	564

The mailings of sets of materials to the survey liaisons and the packets of surveys for the potential respondents included the same items as the Texas mailing described above (see Appendices C and D for samples), with the exception of the Texas Education Agency letter of encouragement to participate; in Illinois, no letter of encouragement was included.

Of the 564 survey packets sent to the survey liaisons in Illinois, 297 usable surveys (52.7 percent) were returned to the researcher.

Massachusetts Sample

In Massachusetts, the researcher used local sources to identify potential survey liaisons in a variety of towns and cities in the western part of the state; Boston was contacted but was unable to participate. In selecting potential survey liaisons, the diversity of the desired respondent sample was a central consideration. A list of 14 potential liaisons was compiled, each of which the researcher contacted by telephone. In the end, 12 educators from 11 different municipalities agreed to serve as survey liaisons in their schools or school districts. Each one was sent the number of surveys that he or she felt could be distributed, plus about 10 percent to cover lost or damaged surveys.

The list of municipalities in the Massachusetts sample and the numbers of survey packets sent in that state is presented in Table 3.3.

Table 3.3
Survey Packets Sent to Massachusetts

Municipality	Number Sent
Amherst	100
Conway	12
Greenfield	36
Hadley	24
Lanesboro	15
Pittsfield	12
Richmond	10
South Deerfield	50
Springfield	50
Sunderland	24
Whately	12
Total	345

The mailings of sets of materials to the survey liaisons and the packets of surveys for the potential respondents included the same items as the Texas mailing described above (see Appendices C and D for samples), with the exception of the Texas Education Agency letter of encouragement to participate; in Massachusetts, a letter of encouragement from the dean of the University of Massachusetts School of Education was included (Appendix E).

Of the 345 survey packets sent to the survey liaisons in Massachusetts, 143 usable surveys (41.4 percent) were returned to the researcher. In all, of the 1483 survey packets sent out to the survey liaisons in three states, 671 usable surveys (45.2 percent) were returned to the researcher.

Data Analysis

Once the questionnaires were returned to the researcher, any enclosed request forms for executive summaries of results were separated from the surveys, as was any extraneous information that might serve to identify individuals or individual schools (e.g., notes to the liaisons, return envelopes with addresses). Then the questionnaires were prepared for tabulation. Since the respondents had been instructed to respond in their questionnaire booklets by circling appropriate responses, the data were key entered into computer-readable format.

In entering data, completely blank surveys were not entered at all; they are not part of the sample of 671 respondents. For individual questions, blank responses were coded as missing and ambiguous responses either were coded as missing (in the cases where two or more contradictory responses such as "Agree" and "Disagree" were circled) or were assigned to one of two complementary categories (i.e., if respondents circled both "Disagree" and "Strongly Disagree," their responses were recorded alternately as one or the other). In addition, a decision was made to consider all "no opinion" responses as missing data.

Questionnaires with handwritten comments were separated from the others for later analysis.

The resulting data set was used with the SAS System of statistical analyses. A variety of analysis procedures was used, depending on the research question being explored. To ensure that the distribution of survey forms (A, B, N, and S) had been random across all three states, a state by survey form chi-square analysis was run, which produced no significant variations from a random distribution. To describe the sample, frequencies of responses to each of the options in the background questions were calculated. The distribution of respondents by ethnicity and gender was compared for Texas and Illinois respondents with data on these important characteristics for the public elementary school

teacher population in these two states (similar data for Massachusetts were not available); results are described in Chapter 4.

For the general purpose questions in Section I of the questionnaire, which all 671 of the respondents had an opportunity to answer, means and standard deviations as well as frequency distributions were calculated. To assess the strength of respondents' opinions regarding the various uses for SATB score reports, the distance of the mean of responses to each question from the midpoint of the scale (3.0) was considered; if the mean differed by more than 0.5 scale points (generally about one half a standard deviation for these respondents) in either direction (i.e., if it was greater than 3.5 or less than 2.5), it was regarded as a strong opinion for this sample of respondents. In addition, as another way to consider strength of opinion, the distribution of responses was examined; any proposed purpose statements for which there was a comparatively high number of "Strongly Agree" or "Strongly Disagree" responses (operationally defined for this sample as 10 percent or more of those who responded) were considered to be of interest, given that these endpoint options had been selected relatively infrequently by the respondents.

To assess whether the variation in the responses to the Section I general purpose questions could be statistically attributed to systematic variations in the background characteristics of the respondent sample, a general linear model (GLM) form of analysis of variance (ANOVA) was run. The GLM form of analysis was selected as appropriate for this survey rather than simple ANOVA because the data in this study were not balanced (i.e., equal numbers of observations were not obtained for every combination of the independent variables in the analysis). In this case, the general linear model is preferable [SAS Institute, Inc., 1990]. Significance was assessed by considering Type III SS estimable functions (i.e., partial sums of squares); the significance level used was $p < 0.05$.

The background variables considered were:

- state
- grade
- teaching setting
- population of municipality of teaching assignment
- teaching experience
- preservice coursework in testing
- inservice coursework in testing
- inservice workshops in testing
- usefulness of testing training
- frequency of use of testing knowledge
- gender
- ethnicity
- familiarity with particular SATBs
- current SATBs used

Differences in responses by survey form were also examined.

For some of the independent variables for which significant F values were reported using GLM analysis, follow-up comparison tests, using the Tukey studentized range test procedure for multiple comparisons, were applied. This procedure is designed to control the error that results from running multiple paired comparison t tests on a number of means (the experimentwise error rate). It is a powerful, conservative test, sometimes called the "honestly significant difference test" [SAS Institute, Inc., 1990]. Comparisons for which the Tukey test revealed no significant differences are not reported in this study as significant, even if significant F values were found through the overall GLM analysis.

For the Section II purpose questions (the first nine questions accompanying each score report sample), only a subset of the overall respondent sample was included in the analysis,

since not every respondent had received the same Section II. Of the 671 respondents, 307 had received the class-level type of score report sample (i.e., Form A or B of the questionnaire) and the remaining 364 had received the individual student type of score report sample (i.e., Form N or S of the questionnaire).

Using these subsamples, similar analyses were run for the purpose questions as for the Section I questions, including means, distributions, GLM ANOVA, and follow-up Tukey comparison tests, to determine the extent of respondent agreement with each purpose question and to examine whether any independent variables were significantly and systematically contributing to the variance in responses. In addition, since each subsample provided opportunities for the same respondents to answer parallel questions regarding two different formats of score report, paired comparison *t* tests were conducted on corresponding questions for each score report format.

For the Section II interpretive questions, the same analyses were conducted as for the Section II purpose questions, including paired comparison *t* tests where parallel questions were available. In addition, mean ratings on each interpretive question were compared with the informal results achieved by the administration of these questions to the local psychometric group (LPG) described in Section 3.3 above.

Because a relationship between respondents' varying amounts of training in testing and measurement issues and the tendency to interpret the sample score reports more or less cautiously had been hypothesized by the researcher, special attention was paid to the three background questions that addressed the amount of training in these areas. The GLM analyses were verified by correlational analysis, using Pearson product-moment methodology. After the GLM analysis and the correlational analysis had revealed no systematic relationship, the responses to these variables were recombined so that instead of levels of training being spread over five response options (i.e., no training, one course, two courses, three courses, and more than three courses), levels of training became a dichotomous variable: no training and some training. Then, GLM analyses were

conducted again. Finally, levels of agreement with the interpretive questions were also collapsed into two categories (disagreement and agreement) and a chi-square analysis was conducted.

CHAPTER 4 RESULTS

Introduction

In this section, the results of the descriptive analyses of the 671 survey respondents will be presented first, based on the Section III background questions, followed by analyses of the responses to the Section I questions, the purpose questions of Section II, and the interpretive questions of Section II. Finally, open-ended respondent comments handwritten on the questionnaire forms will be discussed briefly.

Description of the Respondent Sample

States and Forms

In all, 671 usable (i.e., not completely blank) questionnaires were returned, 231 (34.4%) from Texas, 297 (44.3%) from Illinois, and 143 (21.3%) from Massachusetts. Each respondent completed one of four forms of the survey, representing two different sets of sample score reports. Two sample class-level score reports, in different orders, were contained in Form A and Form B; two sample individual student score reports, in different orders, were contained in Form N and Form S. Each form of the questionnaire took its name from the sample score report that appeared *first* in the form: A (a numerical class-level score report), B (a graphical class-level score report), N (a narrative individual student score report), and S (a numerical/pictorial individual student score report). Forms of the questionnaire were to be distributed randomly among the respondent sample. The result was that 166 respondents completed Form A questionnaires (24.7%), 141 Form B (21.0%), 177 Form N (26.4%), and 187 Form S (27.9%). Thus the total number of respondents who completed either Form A or Form B (referred to here as Form A/B) was 307 (45.8%);

the total number who completed Form N/S was 364 (54.2%). Chi-square analysis of this distribution revealed no significant ($p < 0.05$) difference from an expected distribution.

The distribution of survey forms by state was similarly random according to chi-square analysis. The distribution is displayed in Table 4.1.

Table 4.1
Questionnaires Returned, by State

	TX	IL	MA	Total
Form A	61	73	32	166
Form B	48	59	34	141
Form N	66	74	37	177
Form S	56	91	40	187
Total	231	297	143	671

Grade Levels

Respondents' grade level assignments were most numerous at grades one through six, but the other grades mentioned on the questionnaire were also represented. For this question (Section III, Question 1), respondents were permitted to circle as many grade levels as applied. The result was that 921 grade levels were represented (including "Other") among the 671 respondents. The grade breakdown is displayed in Table 4.2.

Table 4.2
Respondents' Grade Assignments

Grade	N	Grade	N
K	56	5	115
1	103	6	92
2	109	7	74
3	115	8	69
4	124	Other	64

Work Environment

Respondents were mostly teaching in self-contained classrooms, as expected at this level. Out of 664 respondents, 407 (61.3%) reported that their primary work assignment was a self-contained classroom, compared with 188 (28.8%) in departmentalized settings, 8 (1.2%) in multisetting/itinerant assignments, 15 (2.3%) in administrative assignments (e.g., instructional coordinator for math), and 46 (6.9%) in an unspecified other assignment.

Population of Municipality of Teaching Assignment

The plurality of the respondents work in towns with populations between 25,000 and 99,999 people: 231 respondents (37.9% of the 609 who responded to this question) selected this option. Another 139 (22.8%) work in smaller towns (under 25,000 population), while 61 (10%) work in municipalities with populations between 100,000 and 249,999; 46 (7.6%) in municipalities with populations between 250,000 and 499,999; and 132 (21.7%) in cities over 500,000 in population. A substantial number of survey participants (62) did not respond to this question.

The distribution of the respondents to this question by state is worth inspecting. It is summarized in Table 4.3.

Table 4.3
Population of Municipality of Assignment, by State

	Under 25,000	25,000 to 99,999	100,000 to 249,999	250,000 to 499,999	Over 500,000
TX	4	26	34	40	107
IL	82	170	8	2	22
MA	53	35	19	4	3
Total	139	231	61	46	132

As can easily be seen, the Texas sample is more urbanized than samples from the other two states, and the Illinois sample reports particularly small municipality sizes. The

state by size distribution is of course significantly different than expected ($p < 0.001$) according to chi-square analysis.

Years of Teaching Experience

The respondent sample is an experienced group. Only 35 (5.3%) of the 666 persons who responded to this question reported fewer than three years of teaching experience, while 117 (17.6%) reported from three to nine years, 322 (48.3%) from 10 to 20 years, and 192 (28.8%) more than 20 years of experience. Again, the state breakdowns are of interest (in general the Texas sample is less experienced than the Illinois and Massachusetts samples) and the distributions are significantly different than expected ($p < 0.001$) according to chi-square analysis. The distribution is summarized in Table 4.4.

Table 4.4
Years of Teaching Experience, by State

	Under 3	3 to 9	10 to 20	over 20
TX	22	58	100	47
IL	10	34	154	98
MA	3	25	68	47
Total	35	117	322	192

Testing and Measurement Training

The three questions that dealt with respondents' educational experiences in testing and measurement (one focusing on preservice coursework, the second on inservice coursework, and the third on noncredit inservice workshops) yielded response patterns that were generally consistent across states. Responses to the three questions for all respondents are summarized in Table 4.5. In addition to response frequencies in each cell, percents of respondents selecting each response are also given.

Table 4.5
Number of Courses and Workshops and
Percent of Respondents in Each Category

	None	1	2	3	More than 3
Preservice Courses	184 27.8%	262 39.5%	138 20.8%	38 5.7%	41 6.2%
Inservice Courses	361 54.2%	131 19.7%	5 12.8%	27 4.1%	62 9.3%
Workshops	325 48.9%	124 18.7%	88 13.3%	32 4.8%	95 14.3%

Usefulness of Training and Frequency of Use of Testing Knowledge

Questions 8 and 9 of Section III of the questionnaire asked respondents how useful their preparation had been for dealing with testing issues that arise on the job and how often they had to draw on their testing knowledge on the job. Although a majority of respondents to these questions reported finding their preparation at least "sometimes useful" and having to draw on their knowledge at least "occasionally" (defined as 6 to 15 times a year), fully 34.1% of those who responded to the usefulness question selected "not at all useful" or "rarely useful" and 43.1% of those who responded to the frequency question selected "never" or "rarely (1 to 5 times a year)." These results are summarized in Table 4.6.

Table 4.6
Usefulness of Preparation in Testing and
Frequency of Use of Testing Knowledge

Not at all useful	Rarely useful	Sometimes useful	Generally useful	Very useful
113 17.5%	107 16.6%	222 34.4%	161 25.0%	42 6.5%
Never use	Rarely use	Occasion-ally use	Often use	Very often use
14 2.1%	271 40.9%	230 34.7%	81 12.2%	66 10.0%

Gender and Ethnicity

Of the 657 respondents who answered the gender question, 575 (87.5%) were female and 82 (12.5%) were male. Of the 645 respondents who answered the ethnicity question, 3 (0.5%) were American Indian/Alaskan Native; 1 (0.2%) was Asian/Pacific Islander; 92 (14.3%) were Black, Non-Hispanic; 44 (6.8%) were Hispanic; and 505 (78.3%) were White, Non-Hispanic. The state-level breakdowns of these figures are summarized in Table 4.7.

Table 4.7
Gender and Ethnicity, by State

State	Female	Male	Amer. Indian	Asian	Black	Hispanic	White
TX	210 92.5%	17 7.5%	1 0.5%	0 0.0%	19 8.6%	29 13.1%	172 77.8%
IL	248 85.5%	42 14.5%	2 0.7%	1 0.4%	69 24.0%	11 3.8%	205 71.2%
MA	117 83.6%	23 16.4%	0 0.0%	0 0.0%	4 2.9%	4 2.9%	128 94.1%

In the cases of Texas and Illinois it was possible to compare the gender and ethnicity distributions in the sample to recent distributions for the population of public elementary school teachers in those states. In Texas, the Texas Education Agency provided the researcher with figures from the fall 1992 Public Education Information Management System (PEIMS) for the "Total Number of Teachers by Sex and Ethnicity for Grades K-8 from PEIMS Fall 92 Data." In Illinois, the Illinois State Board of Education provided data from its Teacher Service Record for the 1990-91 school year, which lists the number of public school teachers in the state by sex and ethnicity. In Illinois, the data covered both elementary and secondary teachers at the level of the particular assignment, so that the figures for elementary teachers were derived by combining numbers from more detailed breakdowns. Thus the figures reported here are the sum of "elementary education-self

contained," "Chapter I reading and math," "non-Chapter I reading and math," "early childhood," "bilingual education," and "English as a second language" on the Illinois report.

Attempts to gather similar information for Massachusetts were unsuccessful; the Massachusetts Department of Education does not keep figures on public school teachers by gender and ethnicity. The Texas and Illinois percentage figures, together with the corresponding figures from the respondent sample, are summarized in Table 4.8. Beyond the observations that the Texas sample contains a greater proportion of females than the teacher population and that the Illinois sample contains a greater proportion of Black teachers than the teacher population, the sample figures closely resemble the population figures.

Table 4.8
Texas (1992) and Illinois (1990) Teachers' Gender and Ethnicity (%):
Respondent Sample and Public School Teaching Population (K - 8)

State & Group	Female	Male	Amer. Indian	Asian	Black	Hispanic	White
TX Sample	92.5%	7.5%	0.5%	0.0%	8.6%	13.1%	77.8%
TX Population	85.1%	14.9%	0.1%	0.3%	8.5%	14.6%	76.6%
IL Sample	85.5%	14.5%	0.7%	0.4%	24.0%	3.8%	71.2%
IL Population	89.5%	10.5%	0.04%	0.7%	17.0%	3.3%	78.9%

Familiarity With and Use of SATBs

The last two background question on the questionnaire presented respondents with the names of the five major SATBs: the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Test of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), and the Stanford Achievement Test (SAT).

Respondents were asked to indicate the SATBs with which they were familiar and then the SATBs that were currently in use in their schools. Respondents were invited to circle all the SATBs that applied. Table 4.9 summarizes the results.

Table 4.9
SATB Familiarity and Use

	CAT	CTBS	ITBS	MAT	SAT
Familiar	336	279	413	220	271
Use	171	141	279	86	57

Summary of the Sample

In summary, the picture of the respondent sample that emerges from these background data is a group of educators with a good deal of experience, mostly working as teachers in self-contained classrooms and in relatively small to moderate-size towns. They are ethnically diverse, predominantly female, and generally familiar with a variety of SATBs. Their training in testing varies considerably, but most have taken one or two preservice courses that focused on testing and measurement issues and little or no further coursework, either formal or informal, in this area.

The Section I (General Purpose) Questions

Section I of the questionnaire presented 20 statements beginning with the stem "SATB score reports provide useful information for" Respondents were asked to indicate their degree of agreement with each statement on a five-point Likert-type scale ranging from "Strongly Disagree" at scale point 1 to "Strongly Agree" at scale point 5. An "n/o" option was given to those with no opinion regarding a statement. Absent responses and n/o responses were coded as missing data. The scale points were treated as continuous variables and means of responses were calculated: the higher the mean, the greater the mean degree of agreement with the statement. Since every respondent received the same

Section I, regardless of survey form, means are based on a potential maximum of 671 responses.

Mean Ratings

Table 4.10 on the next page presents numbers of responses, overall mean ratings, and standard deviations for each Section I general purpose question. The questions are reproduced in Appendix A and listed in Chapter 3; for ease of understanding they are summarized in capsule form here.

It is reasonable, upon inspection of the data, to regard any opinion as strong for this group that differs from the midpoint of the scale range (i.e., 3.00) by 0.50 or more. This amounts to about one-half a standard deviation in either direction, and points to those statements with mean ratings of 2.50 or less or 3.50 or more. Using this rough metric, it can be said that the teachers in this sample felt strongly *supportive* of the use of SATB score reports for individual student diagnosis (statement 3: 3.73), measuring individual students' growth in broad subject areas (statement 6: 3.59), grouping students temporarily for instruction (statement 2: 3.56), and measuring group achievement in broad areas over time (statement 8: 3.54). The same metric reveals strong *disapproval* of the use of SATB score reports for establishing students' grades (statement 10: 1.72), helping administrators evaluate teachers (statement 20: 1.83), and making promotion/retention decisions for students (statement 11: 2.24).

Frequency Distributions

Inspection of frequency distributions is another way to gauge the comparative strength of respondents' opinions from statement to statement. Table 4.11 summarizes the frequency of responses to each point on the Likert-type scale for each Section I statement.

(SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree.)

Table 4.10
Purposes of SATBs:
Number of Respondents, Mean Ratings, and Standard Deviations

Proposed Use	N	Mean	SD
1. permanent grouping	658	2.65	1.30
2. temporary grouping	665	3.56	1.06
3. individual diagnosis	658	3.73	0.96
4. staying on track with nation	657	3.26	1.06
5. measuring individuals' growth in particular skills	662	3.49	0.99
6. measuring individuals' growth in broad subject areas	664	3.59	0.94
7. measuring group skill achievement over time	661	3.48	0.93
8. measuring group achievement in broad subjects over time	659	3.54	0.93
9. tailoring instruction	660	3.13	1.14
10. establishing student grades	661	1.72	0.99
11. making promotion decisions	659	2.24	1.12
12. evaluating curriculum	660	3.20	1.09
13. self-evaluating effectiveness	665	3.07	1.12
14. helping students know strengths and weaknesses	653	3.03	1.12
15. unexpected teacher insights into students	660	3.23	1.02
16. comparing aptitude and achievement	658	3.26	1.02
17. overcoming bias in teacher judgments	651	2.96	1.06
18. explaining to parents	665	3.47	0.98
19. program comparisons across schools or classrooms	636	2.97	1.11
20. evaluating teachers	648	1.83	1.03

Table 4.11
Purposes of SATBs:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
1. perm. groups	172 26.1%	160 24.3%	86 13.1%	206 31.3%	34 5.2%
2. temp. groups	40 6.0%	86 12.9%	90 13.5%	362 54.4%	87 13.1%
3. indiv. diag.	27 4.1%	55 8.4%	88 13.4%	384 58.4%	104 15.8%
4. natl. pace	44 6.7%	129 19.6%	138 21.0%	303 46.1%	43 6.5%
5. indiv. skill growth	27 4.1%	103 15.6%	106 16.0%	368 55.6%	58 8.8%
6. indiv. subject growth	24 3.6%	79 11.9%	103 15.5%	397 59.8%	61 9.2%
7. group skill growth	25 3.8%	90 13.6%	128 19.4%	381 57.6%	37 5.6%
8. group subject growth	22 3.3%	85 12.9%	119 18.1%	382 58.0%	51 7.7%
9. tailoring instruction	66 10.0%	146 22.1%	135 20.5%	265 40.2%	48 7.3%
10. student grades	367 55.5%	181 27.4%	55 8.3%	49 7.4%	9 1.4%
11. promotion decisions	209 31.7%	214 32.5%	112 17.0%	114 17.3%	10 1.5%
12. evaluating curriculum	65 9.8%	108 16.4%	154 23.3%	295 44.7%	38 5.8%
13. self-evaluation	74 11.1%	139 20.9%	151 22.7%	270 40.6%	31 4.7%
14. student self- knowledge	72 11.0%	152 23.3%	141 21.6%	259 39.7%	29 4.4%
15. insights into students	46 7.0%	119 18.0%	161 24.4%	304 46.1%	30 4.5%
16. aptitude vs. achievement	49 7.4%	108 16.4%	145 22.0%	333 50.6%	23 3.5%

Continued, next page

Table 4.11
continued

17. teacher bias	74 11.4%	141 21.7%	190 29.2%	228 35.0%	18 2.8%
18. explaining to parents	36 5.4%	89 13.4%	103 15.5%	400 60.2%	37 5.6%
19. comparing programs	78 12.3%	147 23.1%	150 23.6%	239 37.6%	22 3.5%
20. evaluating teachers	326 50.3%	176 27.2%	82 12.7%	55 8.5%	9 1.4%

Inspection of the distribution of responses reveals that a majority of the respondents to statements 2, 3, 4, 5, 6, 7, 8, 12, 15, 16, and 18 agreed with those statements (i.e., rated them either "Agree" or "Strongly Agree") and a majority of the respondents to statements 1, 10, 11, and 20 disagreed with those statements (i.e., rated them either "Disagree" or "Strongly Disagree"). Statements 9, 13, 14, and 19 attracted majorities to neither the agreement nor the disagreement side of the scale.

In a Likert-type scale, respondents typically select the endpoints of the scale relatively infrequently; in that respect this respondent group is no different from most. Most statements drew strong disagreement or agreement (i.e., scale points 1 and 5 respectively) from only 5 or 6 percent of the respondents. In the context of these statements and responses, any statement to which 10 percent or more of the respondents chose "Strongly Disagree" or "Strongly Agree" deserves mention.

On the agreement side of the scale, only statements 2 and 3 were in this category: statement 2 about temporary grouping drew 87 respondents (13.1%) to the "Strongly Agree" rating and statement 3 about individual diagnosis drew 104 respondents (15.8%). Both of these statements also displayed "strong" means and were discussed above.

On the disagreement side of the scale, more statements engendered strong opinions. Unsurprisingly, the three statements with comparatively low means, discussed above, drew strong disagreement from relatively large numbers of respondents. Statement 10 (student

grading; mean 1.72) induced 367 respondents (55.5%) to disagree strongly, statement 11 (promotion/retention decisions; mean 2.24) drew 209 (31.7%), and statement 20 (evaluating teachers; mean 1.83) drew 326 respondents (50.3%) to the "Strongly Disagree" rating.

Several other statements that did not display particularly low means, however, drew a number of strongly negative responses. Statement 1 (permanent grouping; overall mean 2.65) induced 172 (26.1%) of the respondents to disagree strongly. For statement 9 (tailoring instruction; overall mean 3.13), 66 respondents (10.0%) selected "Strongly Disagree." Statement 13 (evaluating one's own teaching effectiveness; mean 3.07) caused 74 respondents (11.1%) to disagree strongly, and 72 respondents (11.0%) strongly disagreed with statement 14 (helping students understand their own strengths and weaknesses; mean 3.03). There were 74 respondents (11.4%) who strongly disagreed with statement 17 (overcoming bias in teacher judgments; mean 2.96) and 78 (12.3%) who strongly disagreed with statement 19 (comparing programs across classrooms and schools; mean 2.97).

Thus in some cases a moderate mean might be concealing a fairly strong negative opinion among a substantial number of respondents about some of the proposed uses for SATB score reports.

Relationships between Independent Variables and Section I Questions

Application of the general linear model (GLM) analysis of variance to the data, using each Section I question as the dependent variable and all the background variables as independent variables, resulted in few relationships that might be termed systematic. The most consistent interaction was between the variable "state" and nearly all the Section I questions; other relationships appeared more sporadically between certain Section I questions and background variables. These relationships are described below.

State of Assignment

A consistent pattern of responses emerged from the data when examined by state. Massachusetts respondents consistently rated the Section I questions lower than the Illinois and Texas respondents. In 18 of the 20 Section I questions, the Massachusetts respondents' mean ratings were significantly ($p < 0.05$) lower than the mean Illinois ratings, the mean Texas ratings, or both. A significant difference on these 18 questions was achieved in the overall GLM analysis and in the more conservative follow-up Tukey analysis. In fact only a floor effect in the overall data prevented the two questions (10 and 20) that did not display significant differences from doing so. On the other hand, in none of the 20 Section I questions did the Illinois and Texas ratings differ significantly from one another. Mean ratings by state and for all respondents, with significant differences indicated, are displayed in Table 4.12.

Grade Level

Grade level of teaching assignment influenced the ratings to some of the Section I questions significantly. Notably the teachers in the lower grades were the ones who most often differed significantly from the mean. The following analysis reports every difference that was found to be significant ($p < 0.05$) based on grade level.

Kindergarten teachers tended to disagree with their colleagues on two questions. They supported purpose 18 (explaining to parents) less strongly (3.16) than their colleagues (3.47) and for purpose 16 (comparing aptitude and achievement) they contradicted their colleagues' agreement (2.80 instead of 3.26). Teachers in grade 1 registered four significant differences from the other respondents, in all cases providing lower ratings than the overall mean. With two questions (2, temporary grouping; 3, individual diagnosis) they agreed less strongly (3.32 and 3.54 respectively) than the other respondents (3.56 and 3.73 respectively). Two other questions (19, program comparisons across schools or

Table 4.12
Mean Ratings for Section I Questions, by State

Proposed Use	All	TX	IL	MA
1. permanent grouping	2.65	2.71*	2.87*	2.09
2. temporary grouping	3.56	3.68*	3.66*	3.14
3. individual diagnosis	3.73	3.81*	3.76	3.55
4. staying on track with nation	3.26	3.28*	3.40*	2.95
5. measuring individuals' growth in particular skills	3.49	3.52*	3.61*	3.20
6. measuring individuals' growth in broad subject areas	3.59	3.61*	3.74*	3.25
7. measuring group skill achievement over time	3.48	3.45	3.62*	3.21
8. measuring group achievement in broad subjects over time	3.54	3.48	3.69*	3.31
9. tailoring instruction	3.13	3.22*	3.21*	2.78
10. establishing student grades	1.72	1.67	1.75	1.71
11. making promotion decisions	2.24	2.44*	2.23*	1.93
12. evaluating curriculum	3.20	3.17	3.35*	2.93
13. self-evaluating effectiveness	3.07	3.10	3.17*	2.80
14. helping students know strengths and weaknesses	3.03	3.09*	3.19*	2.60
15. unexpected teacher insights into students	3.23	3.22	3.36*	2.97
16. comparing aptitude and achievement	3.26	3.37*	3.31*	3.00
17. overcoming bias in teacher judgments	2.96	3.07*	2.98	2.74
18. explaining to parents	3.47	3.52*	3.57*	3.19
19. program comparisons across schools or classrooms	2.97	3.08*	3.05*	2.59
20. evaluating teachers	1.83	1.80	1.90	1.77

* significantly ($p < 0.05$) different from MA rating

classrooms; 20, evaluating teachers) elicited stronger disagreement: 2.68 vs. the overall mean of 2.97 on statement 19 and 1.64 vs. 1.83 on statement 20.

Grade 2 teachers, on the other hand, were more positive about two of the statements than the overall mean. They rated purpose 17 (overcoming bias in teacher judgments) more highly than their colleagues (3.11 vs. 2.96) and did the same for purpose 12 (evaluating curriculum) (3.40 vs. 3.20). Third grade teachers rated purpose 4 (staying on track with the nation) more highly than the mean (3.51 vs. 3.26).

Grade 7 teachers were substantially more supportive of purpose 1 (permanent grouping) than their colleagues, turning an overall negative mean (2.65) to a highly positive one (3.83). Finally, grade 8 teachers were more in agreement with purpose 14 (helping students know strengths and weaknesses) than their colleagues, 3.54 vs. 3.03.

These few grade-level differences emphasize rather than contradict a generally consistent pattern of responses across the grade levels. The teacher respondents from different grade levels, with these few exceptions, regarded the 20 potential purposes of SATB score reports quite similarly.

Assignment Setting

No systematic interactions emerged between the work environment question and the Section I questions. The few interactions that registered significant differences and that pertained to respondents who placed themselves in multi-setting/itinerant or administrative categories are not reported because of low numbers of respondents in these categories; the one significant difference that involved respondents in the "Other" category is not reported because this category is not interpretable. On statement 4 (staying on track with the nation), self-contained classroom teachers agreed more strongly (3.34) than departmentalized teachers (3.09), but on statement 14 (helping students know strengths and weaknesses) the pattern was reversed: departmentalized teachers rated this statement 3.22 while classroom teachers rated it on the negative side of the scale (2.94).

Population of Municipality of Assignment

The population of the town or city in which the respondents held their primary assignments made little difference in the responses to the Section I questions. Only two questions displayed any significant differences. Although all categories of respondent rated statement 11 (making promotion decisions) and statement 20 (evaluating teachers) on the negative side of the scale, the 231 teachers from towns between 25,000 and 99,999 in population who responded to statement 11 rated it significantly lower (2.03) than the respondents in the other categories, and the 59 respondents from towns between 100,000 and 249,999 in population who responded to statement 20 rated it significantly higher (2.25) than the respondents in the other categories.

Teaching Experience

No significant differences in the responses to the Section I questions were observed across the different categories of teaching experience. Regardless of the number of years of experience, the respondents answered similarly.

Training in Testing and Measurement

For the three background questions that pertained to respondents' training in testing and measurement issues (Section III, questions 5, 6, and 7), significant, systematic interactions appeared for only the question (5) that related to preservice coursework. For that question respondents who had no coursework responded significantly less positively than those with other amounts of coursework to Section I statements 4 (staying on track with the nation), 5 (measuring individuals' growth in particular skills), 6 (measuring individuals' growth in broad subject areas), 7 (measuring group skill achievement over time), 8 (measuring group achievement in broad subjects over time), and 10 (establishing student grades). The significant difference was apparent between the "no coursework" group and both the "1 course" and the "2 course" groups on statements 5, 7, and 8, and between the "no coursework" group and the "one course" group (only) on statements 4 and

6. It should be noted that the differences involved different degrees of *agreement* with these five statements since all groups' mean ratings on these statements were on the agreement side of the scale.

On the other side of the scale, both the "no coursework" group and the "one course" group rated statement 10 significantly lower than the "2 courses" group and the "more than 3 courses" group, although all groups rated this statement on the disagreement side of the scale. Thus there is some evidence that more coursework in testing and measurement issues is correlated with a greater degree of agreement with some of the potential purposes of SATBs.

Usefulness of Testing Preparation

As might have been expected, some Section I questions received significantly higher ratings from those who, in response to background question 8, said they found their training sometimes or very useful for dealing with testing issues on the job than from those who said their training was not at all useful or rarely useful. The Section I statements on which this pattern of differences occurred are statement 3 (individual diagnosis), 5 (measuring individuals' growth in particular skills), 6 (measuring individuals' growth in broad subject areas), 9 (tailoring instruction), 17 (overcoming bias in teacher judgments), and 18 (explaining to parents).

Frequency of Use of Testing Knowledge

It is neither surprising nor particularly informative to learn that those few respondents (14 persons) who reported never drawing upon their knowledge of testing and measurement issues on their jobs rated two of the Section I questions (9: tailoring instruction and 15: unexpected teacher insights into students) significantly lower than the other respondents. More surprisingly, no other significant differences emerged.

Gender

In general, female and male respondents provided similar ratings to the Section I questions. The two exceptions are statement 1 (permanent grouping) and statement 14 (helping students know their strengths and weaknesses). For both of these statements, not only were there differences in degree, but in fact the 82 males' responses were on the agreement side of the scale (3.22 for statement 1 and 3.32 for statement 14) while the 575 females' responses were on the disagreement side (2.56 and 2.98 respectively).

Ethnicity

On seven of the 20 Section I questions, significant differences in mean responses were found for Black respondents and White respondents. In addition, on one of those seven questions, Black respondents and Hispanic respondents also disagreed significantly. The mean responses to these seven questions are summarized in Table 4.13 on the next page for the three ethnic groups in question (B = Black; H = Hispanic; W = White). The total potential number of Black respondents (overall; not necessarily responding to each question) was 92; of Hispanic respondents, 44; and of White respondents, 505.

In all cases of significant differences, Black respondents rated the statements higher than White respondents. In fact, over the 20 statements in Section I, this pattern was uniformly consistent: the Black respondents rated every statement more highly than the White respondents. It is also the case that the Hispanic respondents rated 19 of the 20 statements higher than the White respondents (the sole exception being statement 1). What factors of background, opinion, or school characteristics contribute to the different perceptions of the valid purposes of SATB score reports among Black, Hispanic, and White respondents can only be matters of speculation.

Table 4.13
Mean Ratings of Section I Questions with
Significant ($p < 0.05$) Differences by Ethnicity

Proposed Use	All	B	H	W
3. individual diagnosis	3.73	3.98	3.81	3.69*
5. measuring individuals' growth in particular skills	3.49	3.85	3.56	3.41*
9. tailoring instruction	3.13	3.82	3.35	2.98*
10. establishing student grades	1.72	2.41	1.98	1.54*
14. helping students know strengths and weaknesses	3.03	3.45	3.31	2.93*
19. program comparisons across schools or classrooms	2.97	3.40	3.05	2.87*
20. evaluating teachers	1.83	2.34	1.81*	1.74*

* significantly ($p < 0.05$) different from Black rating

SATB Familiarity and Use

The last two background questions asked respondents with which of five SATBs they were familiar and which were currently in use in their schools. Respondents were free to circle every SATB that applied.

For an unexplained reason, those 171 respondents who reported that the California Achievement Test was in use in their schools rated as many as 12 of the 20 Section I questions significantly more highly than at least one of the other groups using different tests. The 12 statements that showed this pattern were 4 (staying on track with the nation), 7 (measuring group skill achievement over time), 8 (measuring group achievement in broad subjects over time), 9 (tailoring instruction), 10 (establishing student grades), 14 (helping students know their strengths and weaknesses), 15 (unexpected teacher insights into students), 16 (comparing aptitude and achievement), 17 (overcoming bias in teacher judgments), 18 (explaining to parents), 19 (program comparisons across schools or classrooms), and 20 (evaluating teachers).

The Section II Purpose Questions

Section II of the questionnaire is the point at which the sample divides into two groups according to the general type of sample score reports to which respondents were exposed. One group (called A/B) received two sample score reports (Form A and Form B) that provided hypothetical test information about one class of students in grade 5. The other group (called N/S) received two sample score reports (Form N and Form S) that provided hypothetical test information about one student in grade 5.

In the following sections, results pertaining to the Section II purpose questions for the two groups of questionnaire respondents will be presented, first for the A/B group and then for the N/ S group.

Form A/B

Both Form A and Form B contained the same two sample score reports (report A and report B). Score report A, which was placed first in the questionnaire booklet on Form A and second on Form B, presented its class-level data for the math section of the "Hypothetical Skills Achievement Test" (HSAT) in a numerical format. Score report B, which appeared first on Form B and second on Form A, presented its class-level data for the reading/language arts section of the HSAT in a graphical format. A reproduction of a Form A questionnaire in its entirety appears as Appendix A.

On each form the first nine questions in Section II related to the purposes for which the particular sample score report on the facing page of the questionnaire would be useful. In particular, each of the nine statements was introduced by the incomplete sentence "Score Report A [or B] would be useful to a classroom teacher who wants to:" and respondents were asked to indicate their degree of agreement with each statement on the same five-point Likert-type scale as was used in Section I. The nine purpose questions for the two sample score reports were virtually identical across report forms, except that one of them (question

8) referred to the math curriculum on sample A and the reading/language arts curriculum on sample B.

Of the 671 respondents overall, 307 (45.8%) received and responded to Form A/B. In every demographic respect this subsample is a mirror of the overall sample, matching that larger sample in terms of state, grade level, assignment setting, population of assignment municipality, experience, training in testing issues, opinions regarding the usefulness of their testing preparation, frequency of use of testing knowledge on the job, gender, ethnicity, and familiarity with and use of particular SATBs.

For example, in terms of population of the municipality of teaching assignment, the A/B sample contains 62 respondents (22.8%) from towns under 25,000 population (compared with 22.8% in the larger sample), 102 respondents (37.5%) in the 25,000 to 99,999 category (compared with 37.9%), 32 respondents (11.8%) in the 100,000 to 249,999 category (vs. 10.0%), 20 respondents (7.4%) in the 250,000 to 499,999 category (vs. 7.6%), and 56 respondents (20.6%) from municipalities over 500,000 in population (vs. 21.7%). The A/B sample comprises 267 females (89.3%; compared with 87.5% in the larger sample) and 32 males (10.7%; compared with 12.5%). As for ethnicity, the A/B sample contains 39 Black respondents (13.1%; compared with 14.3% in the larger sample), 22 Hispanic respondents (7.4%; compared with 6.8%), and 235 White respondents (79.1%; compared with 78.3%). There was one American Indian/Alaskan Native in the A/B sample and no Asians.

Mean Ratings on Purpose Questions

The mean ratings on the purpose questions for Section II are summarized in Table 4.14. Ratings are given for both score report sample A (the numerical format) and score report sample B (the graphical format). Capsulized versions of the purpose statements are provided in the table; the full statements can be seen in Appendix A. Each statement focuses on the usefulness of the score report to the classroom teacher.

Table 4.14
Section II Purpose Questions for Score Reports A and B:
Number of Respondents, Mean Ratings, and Standard Deviations

Proposed Use	Report	N	Mean	SD
1. compare achievement nationally	A	297	3.94	0.70
	B	298	3.84	0.76
2. understand individual needs	A	296	3.57	0.92
	B	299	3.58	0.90
3. form temporary groups for skill development	A	292	3.68	0.78
	B	296	3.67	0.84
4. identify skills needing emphasis	A	293	3.82	0.70
	B	298	3.80	0.72
5. evaluate teaching effectiveness	A	289	2.83	1.08
	B	296	2.88	1.02
6. target areas where resources or methods need reevaluation	A	293	3.48	0.84
	B	295	3.48	0.82
7. give feedback to parents	A	294	3.67	0.76
	B	298	3.62	0.78
8. evaluate curriculum effectiveness	A	293	3.24*	0.95
	B	293	3.14*	0.97
9. tailor instruction to individual needs	A	287	3.45	0.99
	B	297	3.37	0.95

*Report A and Report B means differ significantly ($p < 0.05$)

Degree of Agreement. If the same criteria for calling a mean degree of agreement or disagreement "strong" are applied to these questions as were applied to the Section I questions (i.e., a strong opinion is a mean of 2.50 or less or 3.50 or more), the respondent sample can be said to have agreed strongly with the same five statements for Form A and Form B: 1, 2, 3, 4, and 7. Moreover, the respondents responded with strong disagreement to none of the nine questions and responded on the negative side of the scale to only one statement: statement 5. This unusual negative response appears to indicate, as was the case in the Section I questions, that the use of SATB score reports for evaluation of the

effectiveness of teaching, even if the evaluation is done by the classroom teacher rather than an external administrator, does not find support among the respondents.

The frequency distributions of the responses are summarized in Table 4.15 for score report A and in Table 4.16 for score report B. (SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree.) Inspection of the distribution of responses for both score report A and score report B reveals a similar pattern: a majority of the respondents to statements 1, 2, 3, 4, 6, 7, and 9 agreed with those statements (i.e., rated them either "Agree" or "Strongly Agree"). As for disagreement, in no case did a majority of the respondents to either the score report A or the score report B purpose statements disagree with a statement. Two statements, 5 and 8, attracted majorities to neither the agreement nor the disagreement side of the scale.

If the same distributional criteria for considering a degree of agreement or disagreement strong are applied to the distribution of responses here as were applied for the Section I questions (i.e., 10 percent or more of the respondents selected an endpoint of the scale), then only one of the statements may be considered to have elicited strong agreement. For statement 1 (compare achievement nationally), 13.1% of the 297 score report sample A respondents selected "Strongly Agree" and 10.1% of the 298 score report sample B respondents did likewise.

Differences in the Numerical (Report A) and Graphical (Report B) Ratings

Because the same group of respondents reacted to the same two score report samples and answered questions for each sample report that focused on the same potential uses for the reports, it is possible to compare the mean ratings given to each Section II purpose question by using a repeated groups *t* test. If the difference between each pair of means is calculated and the null hypothesis (i.e., that the difference between means is itself not significantly different from zero) is rejected, it can be concluded (with 95% confidence) that

Table 4.15
Purposes of Score Report A:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
1. national comparisons	3 1.0%	15 5.1%	19 6.4%	221 74.4%	39 13.1%
2. individual needs	11 3.7%	35 11.8%	44 14.9%	186 62.8%	20 6.8%
3. temporary groups	6 2.1%	24 8.2%	42 14.4%	205 70.2%	15 5.1%
4. skill emphasis	4 1.4%	17 5.8%	27 9.2%	224 76.5%	21 7.2%
5. evaluate teaching	39 13.5%	74 25.6%	79 27.3%	91 31.5%	6 2.1%
6. evaluate methods	6 2.0%	38 13.0%	69 23.5%	169 57.7%	11 3.8%
7. parent feedback	4 1.4%	27 9.2%	45 15.3%	204 69.4%	14 4.8%
8. evaluate curriculum	12 4.1%	58 19.8%	79 27.0%	134 45.7%	10 3.4%
9. tailoring instruction	17 5.9%	37 12.9%	50 17.4%	167 58.2%	16 5.6%

there are significant differences in the mean ratings assigned by the group under the two format conditions (i.e., numerical and graphical). The major factor to which any differences found through this analysis are attributable may be the difference in format across the two report samples. As indicated in Table 4.14, one A – B difference was statistically significant: statement 8 (evaluate curriculum effectiveness) received a higher rating from the 282 respondents who responded to *both* statements when it appeared in connection with the numerical score report than when it appeared in conjunction with the graphical score report. (It should be noted that statement 8 is the only statement that differed in wording between the two report samples, since for report A it addressed the math curriculum that was the subject of score report A and for report B it addressed the

Table 4.16
Purposes of Score Report B:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
1. national comparisons	6 2.0%	19 6.4%	21 7.0%	222 74.5%	30 10.1%
2. individual needs	10 3.3%	36 12.0%	43 14.4%	191 63.9%	19 6.4%
3. temporary groups	7 2.4%	30 10.1%	38 12.8%	201 67.9%	20 6.8%
4. skill emphasis	3 1.0%	23 7.7%	25 8.4%	227 76.2%	20 6.7%
5. evaluate teaching	26 8.8%	91 30.7%	77 26.0%	98 33.1%	4 1.4%
6. evaluate methods	7 2.4%	33 11.2%	76 25.8%	170 57.6%	9 3.1%
7. parent feedback	5 1.7%	32 10.7%	43 14.4%	209 70.1%	9 3.0%
8. evaluate curriculum	13 4.4%	73 24.9%	73 24.9%	127 43.3%	7 2.4%
9. tailoring instruction	12 4.0%	52 17.5%	56 18.9%	167 56.2%	10 3.4%

reading/language arts curriculum that was the subject of score report B; this wording difference should not have affected respondents' ratings since essentially the same judgment was to be made by the respondents regardless of the particular curriculum area in question.)

If this difference pertaining to statement 8 is actual, the array of numbers pertaining to the particulars of the curriculum area covered on the hypothetical test was apparently considered more useful for evaluating the effectiveness of the curriculum than the array of circles with varying amounts of black fill. In general, however, the format of the score report made little difference to respondents in their assessments of the usefulness of report samples A and B for the stated purposes. Both the numerical format and the graphical format were considered generally useful and received approximately equal support.

Relationships between Independent Variables and the Section II Purpose Questions

A general linear model (GLM) analysis of variance, with the Section II purpose questions for Form A/B as dependent variables and the A/B sample background questions as independent variables, yielded little of true significance. Follow-up Tukey multiple comparison tests were applied to those variables that exhibited some pattern of relationship. Statistically significant relationships are reported below.

Of all the independent variables in the model, the one that accounted for the greatest amount of variance in the Section II purpose questions was the state in which the respondent lived. As was the case with the Section I statements, the Massachusetts respondents generally agreed to a lesser degree with the Section II purpose statements than did the Texas and Illinois respondents. In six out of nine statements relating to score report A (numerical format), the Massachusetts mean ratings were significantly lower than the mean ratings from one of the other two states. Curiously, this pattern did not hold true for the statements relating to score report B; no statistically significant differences in mean response across states appeared.

Mean ratings by state for all the Form A/B Section II purpose questions are summarized in Table 4.17. The number of respondents by state for Form A/B was 109 for Texas, 132 for Illinois, and 66 for Massachusetts.

Inspection of these data reveals two interesting points. First, in no case does the Massachusetts difference result in a rating that is on the opposite side of the scale from the Texas and Illinois ratings (i.e., there are no agreement-disagreement splits). Thus, the entire sample tended to agree and disagree with the same statements. Second, for most statements the agreement patterns *across score report formats* are different for Massachusetts on the one hand and for Illinois and Texas on the other. In general, Texas and Illinois respondents rated report A (numerical format) higher than report B (although

Table 4.17
Section II Purpose Questions for Score Reports A and B:
Mean Ratings by State

Proposed Use	Report	TX	IL	MA
1. compare achievement nationally	A	3.97	3.97	3.81
	B	3.89	3.89	3.66
2. understand individual needs	A	3.75*	3.57	3.29
	B	3.64	3.59	3.44
3. form temporary groups for skill development	A	3.83*	3.64	3.52
	B	3.76	3.62	3.60
4. identify skills needing emphasis	A	3.91*	3.85	3.62
	B	3.81	3.82	3.73
5. evaluate teaching effectiveness	A	2.85	2.94	2.59
	B	2.89	2.92	2.75
6. target areas where resources or methods need reevaluation	A	3.49	3.58*	3.26
	B	3.38	3.55	3.50
7. give feedback to parents	A	3.77*	3.70	3.46
	B	3.65	3.65	3.52
8. evaluate curriculum effectiveness	A	3.28	3.31	3.06
	B	3.05	3.22	3.15
9. tailor instruction to individual needs	A	3.64*	3.38	3.25
	B	3.44	3.39	3.23

*significantly different from MA mean rating ($p < 0.05$)

not necessarily with statistical significance) on these statements; the opposite is true for the Massachusetts respondents, who rated almost every statement higher on report B (graphical format).

Other than the interactions on the state variable, little else emerged from the interaction analysis. For example, the 17 respondents with more than three preservice courses in testing and measurement issues rated statement 7 (give feedback to parents) higher on the B report form than the 277 respondents in the other inservice categories. And the 91 respondents with more than 20 years of experience rated the same statement on the B report higher than the 133 respondents with between 10 and 20 years of experience. Such findings, though statistically significant, appear to have no practical or theoretical meaning.

Finally, this set of responses yielded two of the very rare order effects in the study. For both statements relating to providing feedback to parents (question 7 for both reports), those respondents who saw the numerical format *first* rated the feedback question significantly higher than those who saw the graphical format first (3.76 vs. 3.57 for question A7; 3.72 vs. 3.51 for question B7). In both cases the numerical format was rated a bit higher. No conclusions are drawn from this aberrant finding, which most likely constitutes noise in the study.

Form N/S

Both Form N and Form S contained the same two sample score reports (report N and report S). Score report N, which appeared in the questionnaire booklet first on Form N and second on Form S, presented in a largely narrative format test results for an individual student who took the reading and math portions of the "Hypothetical Skills Achievement Test" (HSAT). Score report S, which appeared first on Form S and second on Form N, presented in a numerical/pictorial format the same individual reading and math information from the HSAT. A reproduction of a Form N questionnaire in its entirety appears as Appendix B.

On each form the first nine questions in Section II related to the purposes for which the particular sample score report on the facing page of the questionnaire would be useful. In particular, each of the nine statements was introduced by the incomplete sentence "Score Report N [or S] would be useful to a classroom teacher who wants to:" and respondents were asked to indicate their degree of agreement with each statement on the same five-point Likert-type scale as was used in Section I. The nine purpose questions for the two sample score reports were identical across report forms.

Of the 671 respondents overall, 364 (54.2%) received and responded to Form N/S. As was the case with the A/B subsample, this subsample mirrors the overall sample in every demographic respect, matching that larger sample in terms of state, grade level,

assignment setting, population of assignment municipality, experience, training in testing issues, opinions regarding the usefulness of their testing preparation, frequency of use of testing knowledge on the job, gender, ethnicity, and familiarity with and use of particular SATBs.

For example, in terms of population of the municipality of teaching assignment, the N/S sample contains 77 respondents (22.8%) from towns under 25,000 population (compared with 22.8% in the larger sample), 129 respondents (38.3%) in the 25,000 to 99,999 category (compared with 37.9%), 29 respondents (8.6%) in the 100,000 to 249,999 category (vs. 10.0%), 26 respondents (7.7%) in the 250,000 to 499,999 category (vs. 7.6%), and 76 respondents (22.6%) from municipalities over 500,000 in population (vs. 21.7%). The N/S sample comprises 308 females (86.0%; compared with 87.5% in the larger sample) and 50 males (14.0%; compared with 12.5%). As for ethnicity, the N/S sample contains 53 Black respondents (15.2%; compared with 14.3% in the larger sample), 22 Hispanic respondents (6.3%; compared with 6.8%), and 270 White respondents (77.6%; compared with 78.3%). There were two American Indian/Alaskan Natives in the N/S sample and one Asian.

Mean Ratings on Purpose Questions

The mean ratings on the purpose questions for Section II are summarized in Table 4.18. Ratings are given for score report sample N (the narrative format) and score report sample S (the numerical/pictorial format). Capsulized versions of the purpose statements are provided in the table; the full statements can be seen in Appendix B. Each statement focuses on the usefulness of the score report to the classroom teacher.

Table 4.18
Section II Purpose Questions for Score Reports N and S:
Number of Respondents, Mean Ratings, and Standard Deviations

Proposed Use	Report	N	Mean	SD
1. compare achievement nationally	N	350	3.92	0.72
	S	354	3.90	0.76
2. understand student's strengths and weaknesses	N	348	3.57*	0.93
	S	356	3.84*	0.82
3. create a plan targeted to student's needs	N	347	3.30*	1.01
	S	355	3.60*	0.94
4. learn reading level of student	N	341	3.21*	1.06
	S	351	3.32*	1.03
5. learn student's grade levels in tested skills	N	343	3.61	0.93
	S	355	3.63	0.88
6. help student understand test performance	N	343	3.20*	1.02
	S	350	3.36*	0.98
7. help teacher discuss test performance with parents	N	344	3.67*	0.83
	S	355	3.79*	0.79
8. evaluate strategies, curriculum, resources	N	342	2.79*	1.05
	S	353	2.99*	1.08
9. set up groups to work on skills	N	343	3.16*	1.04
	S	346	3.52*	0.95

*Report N and Report S means differ significantly ($p < 0.05$)

Degree of Agreement. If the same criteria for calling a mean degree of agreement or disagreement "strong" are applied to these questions as were applied to the Section I questions and the Section II purpose questions for Form A/B (i.e., a strong opinion is a mean of 2.50 or less or 3.50 or more), the respondent sample can be said to have agreed strongly with four statements pertaining to report N (statements 1, 2, 5, and 7) and the same four statements pertaining to report S plus an additional two statements (statements 3 and 9). Moreover, the respondents responded with strong disagreement to none of the nine statements regarding either of the two reports and responded on the negative side of the scale to only one statement regarding both reports: statement 8. This unusual negative response appears to indicate, as was the case in the Section I questions and the Section II

purpose questions for Form A/B, that the use of SATB score reports for evaluation of instructional effectiveness, even if done by the classroom teacher rather than an external administrator, does not find support among the respondents.

The frequency distributions of the responses are summarized in Table 4.19 for score report N and in Table 4.20 for score report S. (SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree.) Inspection of the distribution of responses for score report N reveals that a majority of the respondents to statements 1, 2, 3, 4, 5, 7, and 9 agreed with those statements (i.e., rated them either "Agree" or "Strongly Agree"). For score report S, the same statements attracted a majority of the respondents to the agreement categories, as did statement 6. As for disagreement, in no case did a majority of the respondents to either the score report N or the score report S purpose statements disagree with a statement. Two statements pertaining to score report N, 6 and 8, attracted majorities to neither the agreement nor the disagreement side of the scale. For score report S, only one statement, 8, drew a majority to neither side of the scale.

If the same distributional criteria for considering a degree of agreement or disagreement strong are applied here as were applied for the Section I questions and the Section II purpose questions for Form A/B (i.e., 10 percent or more of the respondents selected an endpoint of the scale), then statement 8 as applied to report N is the only instance of strong disagreement, and statement 1 applied to either report N or report S, and statements 2 and 7 applied to report S attracted strong agreement.

Table 4.19
Purposes of Score Report N:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
1. national comparisons	7 2.0%	13 3.7%	24 6.9%	263 75.1%	43 12.3%
2. strengths and weaknesses	11 3.2%	47 13.5%	48 13.8%	215 61.8%	27 7.8%
3. targeted plan	18 5.2%	69 19.9%	68 19.6%	174 50.1%	18 5.2%
4. reading level	21 6.2%	82 24.0%	61 17.9%	159 46.6%	18 5.3%
5. grade levels	13 3.8%	41 12.0%	41 12.0%	221 64.4%	27 7.9%
6. student understanding	17 5.0%	84 24.5%	74 21.6%	151 44.0%	17 5.0%
7. parent feedback	9 2.6%	27 7.8%	60 17.4%	221 64.2%	27 7.8%
8. evaluate curriculum	36 10.5%	115 33.6%	85 24.9%	98 28.7%	8 2.3%
9. forming groups	22 6.4%	83 24.2%	66 19.2%	161 46.9%	11 3.2%

Differences in the Narrative (Report N) and Numerical/Pictorial (Report S) Ratings

Because the same group of respondents reacted to the same two score report samples and answered questions for each sample report that focused on the same potential uses for the reports, the mean ratings given to each Section II purpose question were compared by using a repeated groups *t* test, as they were for the A/B group. The major factor to which any differences found through this analysis are attributable may be the difference in format across the two report samples.

The report N and report S differences are numerous and unidirectional. Seven of the nine statements, as indicated in Table 4.18, received stronger support to a statistically significant degree when applied to report S (numerical/pictorial format) than when applied

Table 4.20
Purposes of Score Report S:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
1. national comparisons	10 2.8%	14 4.0%	19 5.4%	270 76.3%	41 11.6%
2. strengths and weaknesses	9 2.5%	20 5.6%	38 10.7%	241 67.7%	48 13.5%
3. targeted plan	15 4.2%	35 9.9%	60 16.9%	211 59.4%	34 9.6%
4. reading level	18 5.1%	73 20.8%	58 16.5%	182 51.9%	20 5.7%
5. grade levels	10 2.8%	38 10.7%	56 15.8%	222 62.5%	29 8.2%
6. student understanding	17 4.9%	57 16.3%	76 21.7%	182 52.0%	18 5.1%
7. parent feedback	7 2.0%	23 6.5%	43 12.1%	246 69.3%	36 10.1%
8. evaluate curriculum	33 9.3%	95 26.9%	77 21.8%	137 38.8%	11 3.1%
9. forming groups	15 4.3%	45 13.0%	53 15.3%	211 61.0%	22 6.4%

to report N (narrative format). The narrative report format even received significantly lower ratings in relation to the purpose for which narrative reports have been considered especially appropriate and have been recommended by some test publishers: discussing students' test performance with their parents (statement 6). It appears clear that the respondents largely preferred the greater detail of the numerical/pictorial format, although they regarded both formats positively in relation to the proposed purposes in the questionnaire.

Relationships between Independent Variables and the Section II Purpose Questions

A general linear model (GLM) analysis of variance, with the Section II purpose questions for Form N/S as dependent variables and the N/S sample background questions as

independent variables, resulted in a picture generally similar to the results of the corresponding analysis on the Form A/B questions. The major factors of interest in terms of interactions with the ratings on the Section II purpose questions were the state from which the respondents came and respondents' ethnicity. Follow-up Tukey multiple comparison tests were applied to the variables that exhibited some pattern of relationship. Statistically significant relationships that do not appear to be entirely due to random factors are reported below.

State. Of all the independent variables in the model, the one that accounted for the greatest amount of variance in the Section II purpose questions was the state in which the respondent lived. As was the case with the Section I statements and the Section II purpose questions for Form A/B, the Massachusetts respondents generally registered a lesser degree of agreement than the Texas and Illinois respondents. In seven out of nine statements relating to score report N (narrative format), the Massachusetts mean ratings were significantly lower than the mean ratings from one or both of the other two states. And the same was true in six out of nine statements relating to score report S (numerical/pictorial format).

Mean ratings by state for all the Section II purpose questions are summarized in Table 4.21. The number of respondents by state for Form N/S was 122 for Texas, 165 for Illinois, and 77 for Massachusetts.

Inspection of these data reveals several interesting points. First, the preference of the respondents for report S over report N is especially pronounced among the Massachusetts sample. Not only is every mean rating among the Massachusetts group higher for report S than for report N, but in four cases the means "cross the line" between the agreement side of the scale and the disagreement side. Furthermore, the Massachusetts respondents are the only group with *any* disagreement ratings for report S. Even the lowest-rated statement (8)

Table 4.21
Section II Purpose Questions for Sample Score Reports N and S:
Mean Ratings by State

Proposed Use	Report	TX	IL	MA
1. compare achievement nationally	N	3.86	3.99	3.85
	S	3.78	3.96	3.95
2. understand student's strengths and weaknesses	N	3.58*	3.75*	3.18
	S	3.78	3.97*	3.64
3. create a plan targeted to student's needs	N	3.32*	3.53*	2.77
	S	3.61	3.74*	3.30
4. learn reading level of student	N	3.22	3.35*	2.86
	S	3.34	3.45*	3.00
5. learn student's grade levels in tested skills	N	3.69	3.66	3.34
	S	3.64	3.71	3.42
6. help student understand test performance	N	3.25*	3.40*	2.66
	S	3.46*	3.45*	3.03
7. help teacher discuss test performance with parents	N	3.65	3.83*	3.34
	S	3.77	3.90	3.60
8. evaluate strategies, curriculum, resources	N	2.85*	2.91*	2.40
	S	3.04	3.12*	2.64
9. set up groups to work on skills	N	3.31*	3.29*	2.68
	S	3.48	3.66*	3.28

*significantly different from MA mean rating ($p < 0.05$)

registers overall as a disagreement only because of the Massachusetts sample. This causes a reconsideration of the earlier conclusion that the respondents do not support the use of SATB score reports for instructional evaluation; in the case of report S, the respondents from Texas and Illinois mildly agreed with the usefulness of the report for evaluating instructional strategies, curriculum, and resources.

Ethnicity. The only other systematic relationship between background variables and the Form N/S Section II purpose questions involved the ethnicity of the respondents. Black and White respondents gave significantly different ratings to several statements and the differences were unidirectional, with the Black respondents rating the statements higher.

Table 4.22 summarizes the Form N/S ratings for the three ethnic categories with adequate numbers of respondents to permit reasonable analysis.

Table 4.22
Section II Purpose Questions for Sample Score Reports N and S:
Mean Ratings by Ethnicity

Proposed Use	Report	B	H	W
1. compare achievement nationally	N	3.84	4.05	3.94
	S	3.82	3.86	3.92
2. understand student's strengths and weaknesses	N	3.78	3.68	3.54
	S	3.98	3.91	3.81
3. create a plan targeted to student's needs	N	3.80*	3.64	3.19*
	S	3.96*	3.91	3.52*
4. learn reading level of student	N	3.65*	3.38	3.12*
	S	3.69*	3.45	3.24*
5. learn student's grade levels in tested skills	N	3.76	4.05	3.55
	S	3.75	3.82	3.58
6. help student understand test performance	N	3.54	3.48	3.12
	S	3.61	3.71	3.30
7. help teacher discuss test performance with parents	N	3.84	3.71	3.65
	S	3.94	3.81	3.77
8. evaluate strategies, curriculum, resources	N	3.34*	3.05	2.68*
	S	3.37	3.09	2.93
9. set up groups to work on skills	N	3.51	3.38	3.10
	S	3.90*	3.73	3.45*

*Black and White mean ratings differ significantly ($p < 0.05$)

Order Effects. Three of the rare order effects in the study emerged from the analysis of the Form N/S purpose questions, all pertaining to questions attached to the numerical/pictorial format score report. Those respondents who saw the narrative format *first* tended to rate statements 3, 6, and 8 under the numerical/pictorial format more highly than those who saw the numerical/pictorial format first. Perhaps these proposed purposes (creating targeted plans, helping students understand test performance, and evaluating

instruction) benefited from the evident increase in the level of detail presented in the numerical/pictorial format compared with the already viewed narrative format.

The Section II Interpretive Questions

In addition to the purpose questions in Section II, the two different types of questionnaires (Form A/B and N/S) also contained questions requiring respondent judgment in the form of interpretations of the sample score reports in the instrument. Following the Section II purpose questions, numbered 1 to 9, Form A/B contained six interpretive questions, numbered 10 to 15, relating to information in report A and six relating to information in report B. The same arrangement held true for Form N/S, with six interpretive questions pertaining to report N and six to report S.

As described in Chapter 3, the interpretive questions were actually statements intended to be overinterpretations of the information in the score reports. The independent judgments of eight local individuals with training and experience in testing and measurement, referred to as the local psychometric group (LPG), largely corroborated the researcher's judgment that the interpretive questions were, to varying degrees, unjustified from a psychometric perspective on the basis of the sample score reports (see discussion in Chapter 3). The expectation was that teachers' judgments on these questions might differ somewhat from psychometricians' judgments based on the different natures and demands of their daily jobs and, importantly, upon differences in the amount and nature of their training in issues relating to testing. It was expected that the differences in judgments, ranging on a continuum from "more teacherlike" to "more psychometric," might be reflected within the respondent sample in terms of an interaction between such judgments and the amount of training in testing issues that respondents had experienced. Thus the training-related background questions (5, 6, and 7, but especially 5, which focused on *preservice* training of the sort generally offered within teacher education programs) were of considerable interest to the research design and the intent of the study.

In the following sections, responses to the interpretive questions will be considered first for Form A/B and then for Form N/S.

Form A/B

The six statements pertaining to sample score report A and the six statements pertaining to sample score report B are related to one another, but not through the neat and ordered one-to-one correspondence that the nine purpose questions exhibited. Statement A10 corresponds to statement B10; both relate to comparing local test scores to national achievement levels. Statements A11, B11, A12, B12, and B13 all relate to a similar concept. Statement A11 is closely related to B13 (both involve assessing overall class knowledge levels) and to B11 (which focuses on overall knowledge gaps in the class). Statement A12 is related to B11 and B12, since all three involve comparing the amount of work the class needs to do in different areas covered by the test. Statement A13 is related to B14; both are inferences regarding the amount of class time spent on various learning activities. Statements A14, A15, and B15 are independent, although the two A statements involve a similar type of judgment: interpreting one student's test performance in context.

Table 4.23 arranges the mean responses to the Form A/B Section II interpretive questions according to their relationships to facilitate comparison across the different report formats. Ratings of the respondent sample to all 12 questions are given, as are ratings of the eight-person local psychometric group.

Table 4.23
Section II Interpretive Questions for Score Reports A and B:
Mean Respondent and Local Psychometric Group Ratings
for Related Questions

Type of Interpretation	Question No.	Respondent Rating	LPG Rating
comparing local and national results	A10	3.84*	3.00
	B10	4.07*	3.14
comparing whole class knowledge in different areas	A11	3.30*	2.38
	B11	3.10*	2.89
comparing whole class knowledge in different areas	A11	3.30*	2.38
	B13	3.45*	2.88
comparing whole class knowledge in different areas	A12	3.51*	2.63
	B11	3.10*	2.89
comparing whole class knowledge in different areas	A12	3.51	2.63
	B12	3.51	2.38
inferring amount of class time spent on various activities	A13	2.63*	1.57
	B14	2.78*	1.88
interpreting one student's test performance in context	A14	3.11	2.88
interpreting one student's test performance in context	A15	3.53	2.63
comparing two students' knowledge	B15	3.89	3.00

*significantly different A and B ratings ($p < 0.05$)

Mean Ratings

The pattern of responses reveals that, in general, the respondent sample was in agreement with the statements. The only statement that elicited disagreement was the strongest inference and least justifiable overinterpretation in the set: inferring from test scores the amount of class time being spent on various activities.

The LPG uniformly provided lower ratings than the respondent sample and generally fell on the disagreement side of the rating scale. The LPG was neutral on two statements (A10 and B15) and moderately in agreement with one (B10).

Response Distribution

The distribution of responses to the 12 interpretive questions on Form A/B is summarized in Table 4.24 on the next page. A majority of respondents to statements A10, A11, A12, A15, B10, B12, B13, and B15 were on the agreement side of the scale, while no statement drew a majority of respondents to the disagreement side of the scale. If the distributional standard used for the purpose questions is applied here (i.e., strong feelings are indicated by the presence of 10 percent or more of the responses in either of the endpoint categories), then respondents felt strongly supportive of statements A10, A12, A15, B10, B13, and B15; and they were in strong disagreement with statement A13.

Comparison Across Report Formats

The comparison of mean ratings across the two report formats can be drawn from Table 4.23 on the previous page. For the A10 – B10 set, respondents were in strong agreement with the interpretation, but more in agreement with the interpretation when presented in conjunction with the graphical format. However, it may not be the case that format truly influenced this difference in ratings, since the information needed to consider these questions was not affected by the different format. The respondent sample was in agreement with all five statements in which whole class knowledge levels and needs were to be assessed (A11, A12, B11, B12, and B13). Their agreement was more moderate (i.e., closer to the disagreement side of the scale) under the B format in two of the four pairs under consideration, more moderate under the A format in one of the four pairs, and identical across formats in the last pair. For the pair of statements that drew disagreement, ratings were lower under the numerical format than under the graphical format.

Table 4.24
Score Report A and B Interpretive Questions:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
A10. local vs. national	4 1.4%	31 10.8%	19 6.6%	185 64.7%	47 16.4%
A11. class knowledge	16 5.5%	68 23.5%	44 15.2%	136 47.1%	25 8.7%
A12. class knowledge	9 3.1%	61 21.3%	26 9.1%	156 54.5%	34 11.9%
A13. time spent in class	27 10.3%	98 37.3%	86 32.7%	49 18.6%	3 1.1%
A14. context interpretation	20 7.0%	77 27.0%	59 20.7%	111 38.9%	18 6.3%
A15. context interpretation	11 3.9%	43 15.2%	62 21.9%	120 42.4%	47 16.6%
B10. local vs. national	5 1.7%	13 4.5%	15 5.1%	184 63.0%	75 25.7%
B11. class knowledge	15 5.4%	86 30.9%	48 17.3%	113 40.6%	16 5.8%
B12. class knowledge	8 2.8%	54 18.8%	31 10.8%	173 60.1%	22 7.6%
B13. class knowledge	18 6.3%	58 20.3%	30 10.5%	137 47.9%	43 15.0%
B14. time spent in class	25 9.2%	95 35.1%	77 28.4%	63 23.2%	11 4.1%
B15. comparing two students	5 1.7%	17 5.8%	26 8.9%	201 69.1%	42 14.4%

The two context questions (A14 and A15) were not presented in both formats, but in neither case did respondents apparently find reason in the other test scores of the examinees in question to distrust the targeted test scores. Finally, the statement that involved comparing two students' knowledge drew strong agreement from the respondent sample and neutrality from the LPG. It may be the case that judgments on this statement, presented as it was in conjunction with the graphical depiction of contrasting patterns of fully darkened circles, were influenced by that strong graphical pattern to the point of distortion. The

more detailed and less absolute information that might have been concealed beneath the image of a darkened circle was generally not considered.

From this analysis of the overall means and the comparative means of the A/B respondent sample under two conditions of report format, no conclusions about the influence of numerical vs. graphical formats on the nature and accuracy of score report interpretations can be drawn, except that no such influence is apparent.

Relationships between Independent Variables and the Interpretive Questions

Application of GLM ANOVA, with follow-up Tukey analyses, to the A/B interpretive questions, using all background variables in the analysis, revealed no significant relationships, with the exception of one difference in ratings (question A15, which involves comparing two students) between Illinois respondents and Massachusetts respondents. Despite the lack of significant interactions, two tables are presented below for the sake of consistency with previous analyses. Table 4.25 summarizes mean ratings to the A/B interpretive questions by state and Table 4.26 summarizes ratings provided by the three ethnic groups with the largest numbers of respondents.

State Comparisons. The Massachusetts respondents continued their pattern of tending to rate the statements in the questionnaire lower than the respondents from the other states. In the case of the A/B interpretive questions, however, the differences are not pronounced. The fact that the Massachusetts respondents placed four of the 12 statements on the disagreement side of the scale may be a continuation of their pattern of generally lower ratings rather than reflecting a greater ability to reject overinterpreted score report information.

Table 4.25
Section II Interpretive Questions for Score Reports A and B:
Mean Ratings by State

Statement No. and Type of Interpretation	All	TX	IL	MA
A10 comparing local and national results	3.84	3.91	3.87	3.65
A11 comparing whole class knowledge in different areas	3.30	3.47	3.20	3.22
A12 comparing whole class knowledge in different areas	3.51	3.62	3.45	3.43
A13 inferring amount of class time spent on various activities	2.63	2.63	2.62	2.65
A14 interpreting one student's test performance in context	3.11	3.29	3.07	2.87
A15 interpreting one student's test performance in context	3.53	3.59	3.63*	3.22*
B10 comparing local and national results	4.07	4.01	4.17	3.95
B11 comparing whole class knowledge in different areas	3.10	3.13	3.15	2.97
B12 comparing whole class knowledge in different areas	3.51	3.58	3.49	3.44
B13 comparing whole class knowledge in different areas	3.45	3.54	3.47	3.26
B14 inferring amount of class time spent on various activities	2.78	2.90	2.78	2.58
B15 comparing two students' knowledge	3.89	3.91	3.93	3.75

*significant difference ($p < 0.05$)

Table 4.26
Section II Interpretive Questions for Score Reports A and B:
Mean Ratings by Ethnicity

Statement No. and Type of Interpretation	All	B	H	W
A10 comparing local and national results	3.84	3.76	3.86	3.86
A11 comparing whole class knowledge in different areas	3.30	3.46	2.86	3.31
A12 comparing whole class knowledge in different areas	3.51	3.81	3.09	3.51
A13 inferring amount of class time spent on various activities	2.63	2.64	2.50	2.65
A14 interpreting one student's test performance in context	3.11	3.13	3.50	3.04
A15 interpreting one student's test performance in context	3.53	3.50	3.82	3.52
B10 comparing local and national results	4.07	3.89	4.14	4.11
B11 comparing whole class knowledge in different areas	3.10	3.11	2.90	3.13
B12 comparing whole class knowledge in different areas	3.51	3.57	3.82	3.46
B13 comparing whole class knowledge in different areas	3.45	3.22	3.45	3.50
B14 inferring amount of class time spent on various activities	2.78	2.86	3.00	2.75
B15 comparing two students' knowledge	3.89	3.64	4.05	3.93

No significant differences ($p < 0.05$)

Ethnicity Comparisons. There are no statistically significant differences by ethnicity in these responses. The pattern of responses is well mixed.

Order Effects. One of the very few order effects is present in the results for this part of the study. Respondents who saw the graphical format of the reports *first* rated statement

A13 (inferring time spent on class activities from the information on the numerical report) significantly more highly (2.85) than those who saw the numerical format first (2.44). Since both order groups disagreed with both of the inferential statements under both formats, this does not appear to be a very meaningful result.

Training Interactions. Because a hypothesized relationship between the amount of training in testing issues undertaken by respondents and their ability to apply appropriate judgments to the interpretive questions was of research interest in this study, particular attention was paid to examining any possible relationships between the three training-related background questions (relating to preservice courses, inservice courses, and workshops in testing) and the interpretive questions. The GLM procedure, with the Tukey multiple comparisons follow-up, yielded for the A/B sample and interpretive questions no significant interactions at all. This lack of a relationship was confirmed through a simple Pearson product-moment correlational analysis. This yielded nonsignificant correlations on the order of 0.03 or so.

Next the *differences* between the paired sets of interpretive variables were examined to see if a relationship with the training variables could be discerned. GLM analysis yielded nothing.

To ensure that a possible relationship was not being concealed because there were too many levels of training in each training variable (i.e., each one contained five options: no courses, 1 course, 2 courses, 3 courses, and more than 3 courses), the training variables were collapsed into two categories, yielding a dichotomous variable: no courses (old category 1) and some courses (old categories 2 through 5). GLM analysis was applied to these dichotomous variables and the interpretive questions; no relationship was found.

Finally, the interpretive data were also collapsed into two categories: disagreement (old categories "Strongly Disagree," "Disagree," and "Neutral") and agreement ("Agree" and "Strongly Agree"). Neutral ratings were combined with disagreement ratings because of anecdotal evidence from the LPG and written comments on the questionnaires that some

respondents who judged that there was insufficient information on the score report to support the interpretive statements decided to select "Neutral" instead of disagreeing. The questionnaire had been designed to lead to the disagreement categories for this type of situation, because the set of interpretive questions was introduced by a common stem ("On the basis of Score Report [X], it is justifiable to conclude that . . .") intended to guide respondents to disagreement if they could find no evidence to justify a statement. However, there is some evidence that some respondents (and some LPG members) ignored or misinterpreted the stem and selected "Neutral" if they could find no evidence in the score report to justify a statement.

The two sets of dichotomous variables that resulted from these manipulations were subjected to a 2 X 2 chi-square analysis. No significant divergences from an expected random distribution were found.

On the basis of all these analyses it appears safe to conclude that the amount of preservice, inservice, or workshop training in testing issues that respondents experienced had in general no effect on the nature or quality of their responses to the A/B interpretive questions.

Form N/S

The six statements pertaining to sample score report N and the six statements pertaining to sample score report S are related to one another in the following way. Statement N10 corresponds to S10; both relate to interpretations of grade equivalent scores. Statements N11 and S12 correspond; both concern the use of numerical and confidence band information to compare students to a national sample. Statement N12 corresponds to S11; both concern the interpretation of national percentiles. Statements N13 and S14 ask for judgments relating to the summary information at the top of the score reports, and statements N14 and S13 ask for judgments relating to the more detailed information at the

bottom of the score reports. Finally, statements N15 and S15 concern the assignment of students to mastery categories.

Table 4.27 arranges the Form N/S Section II interpretive questions according to their relationships to facilitate comparison across the different report formats.

Table 4.27
Section II Interpretive Questions for Score Reports N and S:
Mean Respondent and Local Psychometric Group Ratings
for Related Questions

Type of Interpretation	Question No.	Respondent Rating	LPG Rating
interpreting grade equivalent scores	N10	3.49*	2.63
	S10	3.69*	2.63
comparing to national results, using confidence bands	N11	2.88*	2.50
	S12	3.44*	2.88
interpreting national percentiles	N12	3.80	4.13
	S11	3.81	3.14
interpreting summary information	N13	3.61*	2.88
	S14	3.37*	1.88
interpreting detailed information	N14	3.35*	3.00
	S13	3.52*	2.63
assigning to mastery categories	N15	3.35*	4.13
	S15	3.02*	2.50

*significantly different N and S ratings ($p < 0.05$)

Mean Ratings

As was the case with the A/B sample, the N/S respondents generally agreed with the interpretive statements. The only one that drew disagreement was N11, in which respondents were asked to consider apparently conflicting information (a national percentile of 53, a range of 48–68, a confidence band that crosses the 50th percentile, and a grade equivalent score of 4.8 in grade 5) and determine whether to agree that the student was above average. They tended to disagree.

In general the LPG ratings tended toward disagreement, most emphatically with S14, in which the confidence bands (but not the other pieces of information) were identical for two areas of the test. However, the LPG agreed with three statements and were neutral about another. LPG ratings were also generally lower than those provided by the respondent sample for the interpretive questions, but this was not the case for N12 and N15.

Response Distribution

The distribution of responses to the 12 interpretive questions on Form N/S is summarized in Table 4.28 on the next page. A majority of respondents to statements N10, N12, N13, N14, N15, S10, S11, S12, S13, and S14 were on the agreement side of the scale, and statement N11 is unique for being the only interpretive statement on either form of the questionnaire to draw a majority of respondents to the disagreement side of the scale. If the distributional standard used for the purpose questions is applied here (i.e., strong feelings are indicated by the presence of 10 percent or more of the responses in either of the endpoint categories), then respondents felt strongly supportive of statements N10, N12, N13, S10, and S11; no statement for reports N and S drew as many as 10 percent of the respondents to the "Strongly Disagree" scale point.

Comparison Across Report Formats

The comparison of mean ratings across the two report formats can be drawn from Table 4.27 on the previous page. For the interpretation of grade equivalent scores, the respondents' ratings were higher when the statement was presented in conjunction with the numerical/pictorial format; however, this may not be due to a format difference since the grade equivalent information was presented identically for the two reports at the top of the form, before the report formats diverged. Significantly higher ratings (i.e., more in agreement with overinterpretations) were also elicited by the numerical/pictorial format for

Table 4.28
Score Report N and S Interpretive Questions:
Distribution (Number and Percent) of Responses

Statement	SD	D	N	A	SA
N10. grade equivalents	20 5.8%	62 17.9%	33 9.5%	192 55.3%	40 11.5%
N11. national bands/ranges	28 8.1%	147 42.6%	23 6.7%	133 38.6%	14 4.1%
N12. national percentiles	11 3.2%	29 8.4%	18 5.2%	246 71.3%	41 11.9%
N13. summary information	12 3.4%	45 12.9%	44 12.6%	211 60.6%	36 10.3%
N14. detailed information	24 7.5%	48 15.0%	61 19.1%	164 51.4%	22 6.9%
N15. mastery categories	22 7.1%	50 16.1%	59 19.0%	158 50.8%	22 7.1%
S10. grade equivalents	11 3.1%	43 12.1%	36 10.2%	215 60.7%	49 13.8%
S11. national percentiles	14 4.0%	30 8.5%	15 4.3%	253 72.1%	39 11.1%
S12. national bands/ranges	16 4.6%	61 17.6%	25 7.2%	216 62.2%	29 8.4%
S13. detailed information	11 3.2%	52 14.9%	51 14.6%	203 58.2%	32 9.2%
S14. summary information	14 4.0%	76 21.9%	58 16.7%	173 49.9%	26 7.5%
S15. mastery categories	18 5.1%	123 35.1%	52 14.9%	144 41.1%	13 3.7%

statements relating to the comparison of national and local results through confidence bands (S12) and to the interpretation of detailed, objective-level information (S13). On the other hand, the narrative format drew higher ratings for the interpretation of summary information (N13) and the assignment of students to mastery categories (N15). In these cases the finer detail of the S format and the pictorial confidence bands at the objective level may have helped prevent some agreement with the statements.

No consistent pattern of more or less conservative interpretations was associated with either report format. Therefore, as was the case with the A/B sample, no relationship between the report formats studied and the nature and accuracy of score report interpretations was found.

Relationships between Independent Variables and the Interpretive Questions

Application of GLM ANOVA, with follow-up Tukey analyses, to the N/S interpretive questions, using all background variables in the analysis, revealed a few significant differences in response patterns among subgroups of respondents. The first variable of interest is state of assignment. Table 4.29 summarizes the mean ratings for the N/S interpretive questions by state.

State Comparisons. The pattern of generally lower ratings from the Massachusetts respondents is maintained in these results. However, ratings for all respondents, even those from Massachusetts, are generally in agreement with the N/S interpretive statements. The Massachusetts respondents found only one further statement than those identified by the Texas and Illinois respondents with which to disagree: S15, pertaining to the assignment of mastery categories. It is again likely, as was speculated regarding the A/B interpretive questions, that it is the Massachusetts respondents' habit of rating the statements low, rather than any greater ability to reject overinterpretations, that produced the differences apparent in the table.

Ethnicity Comparisons. As with the interpretive questions for the A/B sample, no significant differences among the ethnic groups in the N/S sample emerged from the GLM analysis. Results are summarized in Table 4.30.

Table 4.29
Section II Interpretive Questions for Score Reports N and S:
Mean Ratings by State

Statement No. and Type of Interpretation	All	TX	IL	MA
N10 interpreting grade equivalent scores	3.49	3.59	3.51	3.30
N11 comparing to national results, using confidence bands	2.88	2.81	3.00	2.72
N12 interpreting national percentiles	3.80	3.68	3.91	3.75
N13 interpreting summary information	3.61	3.71*	3.66	3.38
N14 interpreting detailed information	3.35	3.36	3.44	3.15
N15 assigning to mastery categories	3.35	3.29	3.47*	3.18
S10 interpreting grade equivalent scores	3.69	3.74	3.73	3.58
S11 interpreting national percentiles	3.81	3.75	3.77	3.82
S12 comparing to national results, using confidence bands	3.44	3.51	3.51	3.56
S13 interpreting detailed information	3.52	3.62	3.55	3.44
S14 interpreting summary information	3.37	3.43*	3.39	3.14
S15 assigning to mastery categories	3.02	3.12*	3.07	2.80

*significantly different ($p < 0.05$) from MA ratings

Table 4.30
Section II Interpretive Questions for Score Reports N and S:
Mean Ratings by Ethnicity

Statement No. and Type of Interpretation	All	B	H	W
N10 interpreting grade equivalent scores	3.49	3.78	3.36	3.42
N11 comparing to national results, using confidence bands	2.88	2.90	2.77	2.89
N12 interpreting national percentiles	3.80	3.77	3.57	3.82
N13 interpreting summary information	3.61	3.82	3.91	3.55
N14 interpreting detailed information	3.35	3.45	3.56	3.30
N15 assigning to mastery categories	3.35	3.39	3.25	3.34
S10 interpreting grade equivalent scores	3.69	3.82	3.86	3.65
S11 interpreting national percentiles	3.81	3.62	3.68	3.82
S12 comparing to national results, using confidence bands	3.44	3.53	3.18	3.52
S13 interpreting detailed information	3.52	3.44	3.90	3.54
S14 interpreting summary information	3.37	3.24	3.86	3.30
S15 assigning to mastery categories	3.02	3.16	3.19	2.96

No significant differences ($p < 0.05$)

Training Interactions. Application of the GLM ANOVA procedure, with the Tukey follow-up analysis, revealed a few significant differences within the training variables. The 73 respondents with two preservice courses in testing rated statement S15 significantly higher (3.38) than those with one preservice course (2.83). However, this comparison does

not form part of a consistent pattern: the other categories of preservice training gave this statement mixed ratings (3.14 from those with no courses, 2.80 from those with three courses, and 2.87 from those with more than three courses). This is a nonsystematic, random pattern.

Then, the 14 respondents with three inservice courses in testing rated several interpretive questions higher than respondents in other categories. They rated statement N13 (interpreting summary information) significantly lower (2.93) than those with one or two courses (3.73 and 3.80, respectively); they rated statement N14 (interpreting detailed information) significantly lower (2.58) than those with more than three courses (3.64); and they rated statement N15 (assigning to mastery categories) significantly lower (2.67) than those with two courses (3.70). In fact these responses at least construct a reasonable pattern: the respondents with three courses gave lower ratings than their colleagues to just these three questions. They were not consistently low raters, nor were their low ratings mirrored by any other category. However, the fact that these respondents gave lower (i.e., "wiser") ratings to these statements than respondents with both less training and more training destroys the meaning of the pattern. The results are inconclusive; they almost certainly relate to unexplained characteristics of these 14 individuals rather than to factors in their training.

The other significant differences were found in the workshop variable. Here the group to watch is the 51 respondents with more than three workshops on testing. They rated statement N10 (interpreting grade equivalent scores) significantly lower (3.15) than the 22 respondents with three workshops (3.91). They rated N13 and S12 (interpreting summary information; comparing to national results, using confidence bands) significantly lower (3.27 and 3.02 respectively) than the 64 respondents with one workshop (3.83 and 3.73). And they rated the same statement S12 significantly lower than the 177 respondents with no workshops (3.62).

Perhaps it is meaningful that most of the significant training comparisons reported in this section pertain to statements keyed to the narrative score report. Greater training in testing might have contributed in a slight and sporadic way to a tendency to reject overinterpretations based on that type of report. However, this pattern—if pattern it be—still does not add up to the systematic relationship initially hypothesized by the researcher.

Correlational analysis applied to the N/S interpretive questions partially supported the findings of the GLM-Tukey analysis above. Slight, but statistically significant, negative correlations were found between the workshop variable and N13 ($r = -0.11$) and S12 ($r = -0.18$); that is, the more workshops attended by respondents, the lower the agreement ratings.

When the *differences* between the paired sets of interpretive variables were examined to see if a relationship with the training variables could be discerned, GLM analysis yielded nothing.

Next, as with the A/B interpretive questions and for the same reason, the training variables were collapsed into two categories, yielding a dichotomous variable: no courses (old category 1) and some courses (old categories 2 through 5). GLM analysis was applied to these dichotomous variables and the interpretive questions; no relationship was found.

Finally, as with the A/B sample and using the same rationale, the interpretive data were collapsed into two categories: disagreement (old categories "Strongly Disagree," "Disagree," and "Neutral") and agreement ("Agree" and "Strongly Agree"). The two sets of dichotomous variables that resulted from these manipulations were subjected to a 2 X 2 chi-square analysis. No significant interactions were found.

On the basis of all these analyses it appears safe to conclude that the amount of preservice, inservice, or workshop training in testing issues that respondents experienced had in general no systematic effect, and at most a small, sporadic effect, on the nature or quality of their responses to the N/S interpretive questions.

Open-Ended Comments

The questionnaires were formatted so as to leave room for respondent comments. Respondents were specifically invited to comment after Section I and at the end of the questionnaire, where space was left for writing. In all, 171 of the 671 usable responses contained some form of written comment. Many of the comments were explanatory notes pertaining to specific questions on the questionnaire (such as those discussed above that led the researcher to infer that some respondents had marked "Neutral" or "No Opinion" as responses to the Section II interpretive questions when they concluded that not enough information had been presented on the score report to justify the given statement). These explanatory comments will not be discussed here.

The handwritten entries that pertained to SATBs in general or to SATB score reports in general will be briefly discussed. In the following discussion, it should be noted that comments, not commenters, are tallied unless the opposite is explicitly stated. That is, there were more comments than commenters, since one respondent may have provided several comments.

As a general indicator of the strength of opinions about the questionnaire and its topic across the states, the total number of comments from each state is of interest. In Texas, 52 questionnaires out of the 231 that were returned (22.5%) contained handwritten comments. In Illinois the number was 61 questionnaires out of the 297 returned (20.5%). In Massachusetts, 58 of the 143 questionnaires returned (40.6%) contained comments. Once again, respondents from Massachusetts appear to differ significantly from their colleagues in their pattern of response.

However, in the tone and import of the comments no differences were observed across states. There was considerable unanimity in the opinions expressed about SATBs and their score reports. Four general groupings of comments are used to describe the comments here, summarized as follows:

- SATBs are not true indicators of students' knowledge;
- SATBs are harmful;
- SATBs are useful under certain conditions; and
- SATB score reporting could be more effective and helpful to teachers.

Not True Indicators. Across the states, 13 comments simply concluded that SATBs were not worth the time and money they take to administer and process. Many further comments elaborated on the major reason cited for this appraisal: that tests are not true indicators of students' knowledge, falling in this regard far short of teachers' assessments. In all, 90 comments focused on this area of criticism.

Many reasons were given for the respondents' opinions that SATBs fail as accurate indicators of student knowledge. Among those cited more than once were a lack of match between the test and the curriculum, the prevalence of guessing, lack of motivation among students to take SATBs seriously, test anxiety, cheating by students and teachers, differential ability to apply test-taking skills, bias in the tests, and the occurrence of "off days" among students.

Harmful. There were 61 comments that cited one or more forms of harm that tests do to students, teachers, and others. Typical comments focused on damage to students' self-esteem; narrowing the curriculum; inducing stress in students, parents, and teachers; taking instructional time from more productive pursuits; and (especially) improperly evaluating the performance of teachers. One set of comments focused on the inappropriateness of such tests at the lower grades (i.e., kindergarten and grade 1).

Conditionally Useful. Respondents provided 39 comments that concluded that SATBs were useful under certain conditions. The conditions most often cited were if SATBs are used as one of many indicators of student performance; if they are used to understand students' needs and to help students; if they are used for initial planning and grouping (i.e., before the teacher knows a lot about the students); if time and effort are expended to understand test results; if group trends rather than individual performances are the focus of

consideration; if tests are limited in number; if they are needed to satisfy political constituencies; and if they are used by the district to make curriculum decisions.

How SATBs Could Be More Useful. The last category of comments (21 comments) comprises suggestions about how SATBs, and especially score information, could be more useful. Suggestions included administering the tests earlier in the year (not in May) so results would be available for current year planning; a shorter lag time between testing and reporting; less complex and ambiguous reports; more help from trained professionals, such as school psychologists; reports going to the teacher instead of to a file in the central office; and results being given at the end of the year to the current teacher instead of to the next year's teacher (so that results could be used for reflection and self-evaluation).

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

General Purposes

Respondents in general registered moderate degrees of support for many of the stated purposes of SATB score reports. The levels of agreement they expressed with the general purpose questions in Section I of the questionnaire cannot be called enthusiastic, but they were certainly not in the negative category, with a few predictable exceptions. This finding should not be regarded as an endorsement of SATBs, but this study did not uncover widespread and intense disaffection with SATBs and their score reports.

Respondents tended to be substantially more supportive of uses of SATB score reports relating to local and provisional decisions that are clearly in teachers' hands and that are subject to reevaluation and revision (such as temporary grouping and student diagnosis) than of uses relating to more permanent and systemwide decisions (such as promotions, grading, and the evaluation of classroom personnel, activities, curricula, and materials). This is a reasonable finding given the nature of the teaching profession. It is certainly unsurprising to find teachers decidedly unsupportive of the use of test scores for teacher evaluation and student grading; Goslin's [1967] study produced similar negativity regarding these purposes.

There appears also to be some support for uses of SATB score reports relating to the gathering of information that is difficult to obtain in other ways (such as student growth over time and, to a lesser degree, national comparisons) and for using SATB score reports as an aid in explaining student performance to parents. There was moderate support of the use of SATB score reports for comparing students' aptitude with their actual achievement and for gaining unexpected insights into students.

For uses pertaining to self-evaluation and to the modification of instructional plans (such as tailoring instruction to individual students' needs and evaluating the curriculum and

curriculum materials) there appears to be little enthusiasm. Even when such evaluative uses were clearly to be confined to the teacher's purview, and were not to be imposed from outside the classroom, they achieved ratings very close to the neutral midpoint of the scale. It is hard to interpret this finding beyond the bare quantitative results, but perhaps SATB score reports provide too little information too late (as some open-ended comments indicated) to be truly useful for this sort of evaluation.

Respondents did not tend to find SATB score reports useful for overcoming potential personal biases regarding students. The impartial, objective nature of the SATB evidently did not figure heavily in respondents' consideration of this potential purpose.

Relationships with Independent Variables. The most striking effect in this portion of the study was the significant difference in response pattern exhibited by the respondents from Massachusetts compared with the other respondents. For 18 of the 20 general purpose questions, the Massachusetts ratings were significantly lower than one or both of the other states' ratings. It seems likely that the lower ratings reflect a lower regard for SATBs in general (and not just for SATB score reports) among the Massachusetts respondents than among the respondents from the other two states. This impression is confirmed by both the number and the content of the Massachusetts respondents' open-ended comments. It is a matter of speculation whether this difference in opinion, as large and consistent as it is, represents a regional difference, a difference in educational traditions or training, or the onset of an emerging opinion shift that may spread to other locations.

The generally lower ratings provided for the general purpose questions by kindergarten and grade 1 teachers are consonant with open-ended comments from several respondents; it was expected that the controversial issue of testing at the lower grades would be reflected in the ratings to this questionnaire. More surprising were the generally higher ratings to the general purpose questions provided by the Black respondents in comparison with the White respondents. It is of interest that, despite considerable criticism

of SATB testing on grounds of bias, one minority group would appear to have higher regard for many of the purposes of SATB score reports than the majority group.

Finally, it is interesting that the amount of preservice training in testing and measurement issues experienced by respondents apparently is related to their opinions on at least some of the general purpose questions (those relating to national comparisons, growth of individuals and groups over time, and grading), especially in light of the finding that training had virtually no effect on other aspects of this study. However, the differences that were found between the "no preservice training" group and other amounts of preservice training did not constitute disparate judgments on these proposed purposes (i.e., there were no agreement-disagreement splits), but merely differences in the strength of respondents' agreement or disagreement, with the "no training" group providing lower ratings.

Score Report-Related Purposes and Preferences

Form A/B. When faced with specific examples of class-level score reports (i.e., reports A and B), respondents' opinions about their potential uses were in general consistent with those expressed through their ratings of the general purposes in Section I, although nearly all ratings were higher and some ratings showed slight differences. The use of both forms of the class-level score report to create temporary groups for skill development was rated high, as was their use for individual diagnosis and for giving parents feedback on their children's skills. The use of the reports for evaluating curriculum effectiveness received similarly modest approval when linked to a specific score report in Section II and in general (in Section I), but the uses of the score reports for tailoring instruction, targeting areas where resources or methods might need reevaluation, and comparing achievement with national levels were rated higher in Section II than in Section I. Perhaps the presence of the score report served as a reminder of the nature and utility of the information that was available to teachers on such instruments.

The respondents' equal regard for the graphical and the numerical format of the class-level score report was a surprising finding to the researcher. The array of numbers virtually covering a full page (score report A) seemed daunting to the researcher, especially when compared with the graphical array of circles with varying amounts of fill (score report B). However, the differences in format had no apparent effect on the opinions of the respondents regarding the purposes for which the sample score reports might be useful.

Form N/S. As with the A/B score reports, the respondents' purpose ratings when faced with actual samples of score reports in Section II were often higher, especially for the numerical/pictorial format, than when score report purposes were considered in the abstract in Section I. Respondents were more positive about use of the individual student score reports for national comparisons, targeting instructional plans, helping students understand their performance, and explaining test results to parents. They were also supportive of learning the student's grade levels in the tested skills, a use not covered in the general purpose questions.

The comparison between the two formats of the individual student score report yielded substantial differences. For seven of the nine statements, respondents' ratings were significantly higher for the numerical/pictorial format (score report S) than for the narrative format (score report N). This result appears to contradict the findings of the Mathews [1972, 1973] studies in which teachers strongly preferred a narrative version of a score report when compared with a more traditional format, presumably one similar to the numerical/pictorial format of this study, for both class-level and individual student information. It may be the case that the Mathews narrative format contained more explicit and detailed information than the version used in this study, or it may be that the task of the teachers in assessing the different score reports in this study was more focused on particular purposes, as opposed to overall impressions, than in the Mathews study. In any event, the contrast in respondents' opinions about the two formats was striking.

Relationships with Independent Variables. The significantly different opinions expressed by the Massachusetts sample in comparison to those expressed by the Illinois and Texas samples is again noteworthy. As was the case with the Section I questions, the difference is notable in both its quantity and its unidirectionality. It is again likely that the differences reflect different opinions regarding SATBs in general, even though they were expressed in conjunction with a questionnaire on score reports. The source of the difference is not known.

Similarly unexplained is the Black–White difference in response pattern for the N/S reports. The Black teachers in the sample appeared to rate more highly than their White colleagues the use of the narrative and the numerical/pictorial format for creating targeted instructional plans and learning the reading levels of students; the use of the narrative format for evaluating strategies, curriculum, and resources; and the use of the numerical/pictorial format for setting up groups of students to work together on skills. Perhaps these differences relate to the higher ratings awarded by the Black respondents to the Section I purpose questions, and reflect a generally more positive opinion regarding SATBs and their score reports.

Interpretive Questions

Teachers' Knowledge of Testing. The findings of this study relative to the 24 interpretive questions generally corroborate the studies reported in the literature review [e.g., Fredrickson & Marchie, 1966; Hills, 1991; Huebner, 1987, 1988]. Teachers' knowledge of testing issues is not as thorough as psychometricians, in general, believe it should be. In the face of score report information, and nothing else, the teachers in this study were generally willing to accept interpretive statements that were at least in large part overinterpretations of that information from a psychometric perspective.

The statements were designed to be typical of the kinds of judgments that teachers are often called upon to make, either formally or informally, about individual students and

classes of students. The statements conveyed imprecise but apparently meaningful judgments of the sort that can affect students' actual learning, study assignments, classroom assignments, self-perceptions, expectations, and attitudes toward schooling. If conveyed to other educators, such statements can influence others' expectations and judgments regarding students. If conveyed to parents, statement of this sort can foster erroneous impressions of their children's efforts and achievements.

It should be noted that the teachers in this study in no way erred only in favor of students in their overinterpretations. It is not the case that the teachers' judgments on the basis of score reports were overly generous; in many cases, their judgments unjustifiably placed individual students at a disadvantage, and in other cases their judgments, if acted upon, would have resulted in instructional decisions that were not necessarily wise for the students.

The consequences of overinterpretations of score reports are therefore serious. It is curious that in all the citations of harmful effects of testing on students offered by the respondents who provided open-ended comments on the questionnaire, none mentioned the harm that overinterpretation of score reports can cause. It is also curious that with all the skepticism that teachers voice about standardized tests, the respondents were so willing to trust rough numerical information as unduly precise, and so unwilling to doubt, to hesitate.

The overinterpretations concerned concepts that are central to the field of testing: concepts of reliability, error, probability, and approximation. And if teachers cannot interpret such concepts ably, the most central psychometric concept of all, validity, becomes at issue [Tittle, 1989]. For as the *AERA, APA, & NCME Standards for Educational and Psychological Testing* [1985] define the concept in the very first paragraph, "[validity] refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" [p. 9]. The inferences from test scores that were presented as part of this study were simply not *valid*.

The Non-Effect of Training. The most significant finding of this study was perhaps not that teachers do not make good "psychometric" judgments about SATB score reports, but that apparently no amount of prior training in testing issues proved adequate to instill appropriate caution in the respondent teachers as they faced the task of considering and analyzing score reports. The amount of testing-related training experienced by the respondents, whether in the form of preservice training, inservice credit courses, or inservice noncredit workshops, exhibited no systematic relationship at all with the accuracy of their interpretations of score report information. Teachers with more testing courses and workshops were evidently as willing as teachers with fewer testing courses and workshops to accept overstated interpretations of score reports; neither group applied caution and skepticism to the task.

Format Differences. Another significant finding of this study was that the variations in score report format used in the questionnaire had no real effect on the accuracy of the teachers' interpretations. Not the graphical format nor the confidence band-based numerical/pictorial format facilitated accurate interpretation, even in the face of the sorts of interpretations for which such formats were designed (i.e., using the graphical format to notice broad areas of need in a classroom or to pick anomalous scores out of an environment with which they are in discord; using the confidence band format to avoid overinterpreting apparent numerical differences or making overly absolute status assignments). Respondents certainly preferred the confidence band-based numerical/pictorial format of the individual student score report, but their preference did not result in improved accuracy in their interpretations.

Recommendations

The following recommendations result from a consideration of the methodology used in this study, the results obtained, and the previous research outlined in the review of the literature.

First, the appropriate conjunction of, on the one hand, teachers' needs in relation to information that can be derived from SATBs and, on the other hand, the contents and formats of the reports that are used to provide such information has not yet taken place and should take place soon. It is clear from this study and others that teachers are not finding the information on score reports genuinely helpful and are not making appropriate and meaningful inferences from such information; the validity of teachers' classroom interpretations based on score reports is therefore dubious. Score reports as they are now presented offer too much information in a format that is not only too complex for ready understanding, but too open to erroneous interpretations.

Teachers should have the major say in the information they will receive on the score reports that they are expected to use; they should also have a say in whether and how to use those reports. The terminology used on score reports should be "teacher-friendly" and *pedagogically* sound (in addition to being psychometrically sound); and teachers should help to work out terminology that they can understand and that is professionally meaningful to them. Technical terms and pseudotechnical terms such as "grade equivalent score" should be avoided rather than presented and feebly bolstered by interpretive manuals that are rarely opened.

Second, teacher training in testing issues should be not increased but strengthened. The focus of training in psychometric matters should be not what psychometricians have traditionally believed teachers ought to know, but what teachers really need to know and want to know. For example, the foundations of testing and measurement theory in mathematical and statistical concepts, however intrinsically valuable and seminal, should be largely ignored except at the conceptual and intuitive levels. Testing classes should involve an action and discovery orientation; teachers and future teachers should become collaborators in their own learning so that testing concepts and methods become meaningful to them as real tools for solving real problems.

The foundation of this teacher focus is the growing realization that teachers and psychometricians inhabit different worlds [Tittle, 1989]. The professional reality of the teacher, and the context within which all pedagogical activities, including the interpretation of SATB score reports, occur is vastly different from the professional reality of the psychometrician. The interpretations that a teacher brings to a score report that a psychometrician regards as a sterling specimen of clarity and directness may be foreign and surprising to that psychometrician, but they may make supreme sense in the context of the teacher's classroom culture. Unless the two realities come together, the incursions of the one on the territory of the other will generally be regarded as either bizarre or hostile.

A final recommendation is that further work be done to clarify the issues raised in this study, which uncovered as many questions as answers. What accounts for the differences in response patterns among teachers from Massachusetts and between Black and White respondents? Why do seemingly skeptical teachers fall prey so easily to overinterpretations of score information? What do teachers really find helpful (and what annoying, neutral, or harmful) in score report contents and formats? What kinds of training, and what sorts of instructional methods, would teachers really find most useful to help them gain as deep an understanding of testing issues as they need?

The answers to these and other questions should be explored not in a further development of survey research, but through a more qualitative methodology that will permit the lengthy and profound exploration of these issues among smaller groups of teachers. A follow-up research study is recommended involving focus group or one-on-one interviewing methodology to build on the findings of this study and illuminate its curious and promising aspects more thoroughly.

APPENDIX A
QUESTIONNAIRE FORM A

**Teacher Questionnaire on the Uses of
Score Reports from Standardized Achievement Test Batteries
(Form A)**

This questionnaire asks teachers for their opinions about the score reports that summarize their students' performance on standardized achievement test batteries. The test batteries of particular interest in this study are the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Test of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), and the Stanford Achievement Test (SAT).

In light of the fact that nearly all school districts administer standardized achievement tests, it is surprising to note that relatively little is known about *teacher* opinions regarding the score reports from these tests. Do teachers find the information in score reports useful? If so, for what purposes? Is the information in these reports understandable? Are the formats of these reports clear and helpful? What changes would teachers like to see? These are the sorts of questions that this study seeks to answer by directly addressing teachers.

You are being asked to respond to this questionnaire because, as an experienced teacher in grades K - 8, you have almost certainly worked with score reports from such tests and formed some opinions about them. Because your opinions are likely to be both informed and practical, they should be of considerable value to those who design and publish tests for classroom use.

We first ask your opinions regarding the potential purposes, in general, of standardized test score reports. Then you are asked to consider two fictitious examples of a widely used type of score report and to answer opinion and interpretive questions about those examples. Finally, we ask some background questions to help describe the sample of teachers who responded to the survey.

In all cases, you are asked to respond on the basis of your own experiences with standardized achievement test score reports and to share your personal opinions and preferences regarding them. Your responses to this questionnaire will be strictly confidential. All information that could identify individual respondents will be eliminated after the questionnaires are returned.

Completion of this questionnaire should take about 30 minutes. We realize that this time expenditure is not trivial in the busy schedule of a teacher, but we hope the knowledge gained on behalf of the education of children will be worth the sacrifice.

If you would like a copy of an executive summary of our findings, please fill out your name and address on the enclosed form. (These forms will be separated from your responses before your responses are tabulated.)

Thank you for your participation.

Ronald K. Hambleton
Professor of Education and Psychology

Edward J. Murphy
Research Associate

University of Massachusetts

Section I: Purposes of Score Reports from Standardized Achievement Test Batteries

Please read the following statements about potential purposes that standardized achievement test battery (SATB) score reports have been said to serve. For each statement, indicate the degree to which you agree that the statement reflects a *valid* purpose that SATB score reports actually serve *in your own experience*. The potential degrees of agreement are as follows:

- SD Strongly Disagree (i.e., the stated purpose is *emphatically not valid* in your experience)
- D Disagree (i.e., the stated purpose is *generally not valid* in your experience)
- N Neutral (i.e., the stated purpose *may or may not be valid* in your experience)
- A Agree (i.e., the stated purpose is *generally valid* in your experience)
- SA Strongly Agree (i.e., the stated purpose is *emphatically valid* in your experience)

If you have no opinion about a statement, circle *n/o*.

<i>Statement</i> (SATB score reports provide useful information for...)	<i>Degree of Agreement</i> (SD = Strongly Disagree to SA = Strongly Agree)	<i>No Opinion</i>
1. helping schools make decisions about placement of individual students into <i>permanent</i> instructional groups (e.g., homogeneous ability groups).	1. SD D N A SA	<i>n/o</i>
2. helping teachers make decisions about placement of individual students into <i>temporary</i> instructional groups (e.g., cooperative learning groups, groups for enrichment or remedial work).	2. SD D N A SA	<i>n/o</i>
3. helping teachers diagnose individual students' strengths, weaknesses, and needs.	3. SD D N A SA	<i>n/o</i>
4. helping teachers keep the pace and level of instruction "on track" with national expectations.	4. SD D N A SA	<i>n/o</i>
5. enabling teachers to measure individual students' growth in particular skills.	5. SD D N A SA	<i>n/o</i>
6. enabling teachers to measure individual students' growth in overall subject areas (e.g., math, language arts).	6. SD D N A SA	<i>n/o</i>
7. enabling teachers to measure group achievement in particular skills over time.	7. SD D N A SA	<i>n/o</i>
8. enabling teachers to measure group achievement in overall subject areas over time.	8. SD D N A SA	<i>n/o</i>
9. enabling teachers to plan instruction that is tailored or adapted to individual students' needs.	9. SD D N A SA	<i>n/o</i>
10. helping teachers establish individual students' grades in class.	10. SD D N A SA	<i>n/o</i>

<i>Statement</i> (SATB score reports provide useful information for...)	<i>Degree of Agreement</i> (SD = Strongly Disagree to SA = Strongly Agree)					<i>No Opinion</i>	
11. enabling schools to make promotion/retention decisions for individual students.	11.	SD	D	N	A	SA	n/o
12. helping teachers and schools evaluate the effectiveness of the curriculum or of curriculum materials.	12.	SD	D	N	A	SA	n/o
13. helping teachers evaluate the effectiveness of their own instructional approaches and strategies.	13.	SD	D	N	A	SA	n/o
14. helping students gain personal insight into their strengths and weaknesses.	14.	SD	D	N	A	SA	n/o
15. helping teachers gain unexpected insights into particular students' hidden talents, achievements, or interests.	15.	SD	D	N	A	SA	n/o
16. enabling teachers to compare individual students' aptitude and achievement levels.	16.	SD	D	N	A	SA	n/o
17. helping teachers eliminate potential sources of personal bias in evaluating their students' abilities by providing an objective form of information on student achievement.	17.	SD	D	N	A	SA	n/o
18. helping teachers explain individual student achievements and needs to parents.	18.	SD	D	N	A	SA	n/o
19. enabling administrators to compare varying programs or approaches being implemented in different classrooms or schools.	19.	SD	D	N	A	SA	n/o
20. helping administrators evaluate the performance of individual teachers.	20.	SD	D	N	A	SA	n/o

If you have any comments, please write them here:

Section II: Sample Score Reports

In this section of the questionnaire you are asked to consider and respond to questions about two sample score reports. These score reports are fictitious; that is, they have been created by the researchers rather than being reproduced exactly from any published test. They are intended, however, to be similar to actual score reports and to present information and format features that are of the sort offered by standardized achievement test publishers.

The two samples are named as follows:

Sample A: Class Diagnostic Analysis Report

Sample B: Class Objective Mastery Report

The two reports provide imaginary test score information for one class of students in the fifth month of fifth grade at the fictitious Tyler School. You may assume that the students took a standardized achievement test battery consisting of several subject area subtests, including reading/language arts and mathematics.

Sample A presents information on the mathematics subtest for all 26 students (arranged alphabetically) in Mr. or Ms. Winston's class. Sample B presents information on the reading/language arts subtest for the same students. Sample A uses a numerical format, while Sample B uses a more graphical format.

Please respond to the questions in this section based on your own personal experience as a teacher.

HSAT Hypothetical Skills Achievement Test

Sample A: Class Diagnostic Analysis Report

DIAGNOSTIC REPORT FOR:

Class: Winston Grade: 5.5
 School: Tyler District: Garfield
 City: Anytown State: Anystate
 Form/Level: B/17 Test Date: 1/20/91
 Norms From: 1988
 National Reference Group: 5.6

NOTE: All scores indicate the percent of test items answered correctly.

	A	B	C	D	E	F	O	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
NUMBER OF STUDENTS:	26	92	54	54	54	38	92	46	92	54	92	38	38	31	46	92	62	54	92	54	92	46	46	92	92	62
MATH CONCEPTS																										
Number Systems/Whole Numbers																										
Fractions																										
Decimals and Percent																										
Equations																										
MATH COMPUTATION																										
Add/Subtract Whole Numbers																										
Multiply/Divide Whole Numbers																										
Add/Subtract Fractions																										
Add/Subtract Decimals																										
Compute Percents																										
MATH PROBLEM SOLVING																										
Single Step Add/Subtract																										
Single Step Multiply/Divide																										
Multistep Problems																										
Critical Thinking Problems																										
MATHEMATICAL EXPRESSION																										
Use of Math Symbols/Terms																										
Interpretation of Graphs/Charts																										

Questions Pertaining to Score Report Sample A: Class Diagnostic Analysis Report

Please consider the statements below in relation to Score Report A on the facing page. Circle the letter in the second column that corresponds to the degree to which you agree with each statement as applied to Score Report A. (SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree.) If you have no opinion about a statement, circle *n/o*.

<i>Statement</i>	<i>Degree of Agreement</i>					<i>No Opinion</i>	
<i>Score Report A would be useful to a classroom teacher who wants to:</i>							
1. compare his or her students' achievement levels to those of students nationwide.	1.	SD	D	N	A	SA	<i>n/o</i>
2. understand the instructional needs of particular students.	2.	SD	D	N	A	SA	<i>n/o</i>
3. form temporary work groups to focus on individual skill development.	3.	SD	D	N	A	SA	<i>n/o</i>
4. distinguish skill areas that need emphasis from those that do not.	4.	SD	D	N	A	SA	<i>n/o</i>
5. evaluate his or her own teaching effectiveness.	5.	SD	D	N	A	SA	<i>n/o</i>
6. target curriculum areas in which resources and/or teaching methods should be reevaluated.	6.	SD	D	N	A	SA	<i>n/o</i>
7. provide feedback to parents on the skills of their children.	7.	SD	D	N	A	SA	<i>n/o</i>
8. evaluate the effectiveness of the math curriculum.	8.	SD	D	N	A	SA	<i>n/o</i>
9. plan instruction that is tailored to individual students' needs.	9.	SD	D	N	A	SA	<i>n/o</i>
<i>On the basis of Score Report A, it is justifiable to conclude that:</i>							
10. compared with students nationwide, this class is below average in "Math Concepts" and above average in "Math Computation."	10.	SD	D	N	A	SA	<i>n/o</i>
11. as a whole, this class knows more about concepts related to "Number Systems/Whole Numbers" than about concepts related to "Decimals and Percent."	11.	SD	D	N	A	SA	<i>n/o</i>
12. of the four areas covered by the test, this class needs the most work in "Math Problem Solving."	12.	SD	D	N	A	SA	<i>n/o</i>
13. this teacher spends too much time on "Math Computation" and not enough time on "Math Problem Solving."	13.	SD	D	N	A	SA	<i>n/o</i>
14. Luisa Ali needs remedial work on "Fractions."	14.	SD	D	N	A	SA	<i>n/o</i>
15. Seth Viola knows as much about concepts relating to "Equations" as Theodora Xavier does.	15.	SD	D	N	A	SA	<i>n/o</i>

HSAT Hypothetical Skills Achievement Test Sample B: Class Objective Mastery Report

OBJECTIVE MASTERY REPORT FOR:
 Class: Winston Grade: 5.5
 Tyler District: Garfield
 Anytown State: Anystate
 Form/Level: B/17 Test Date: 1/20/91
 Norms From: 1988
 National Reference Group: 5.6

○ = Not Mastered (under 50% of items correct)
 ◐ = Partly Mastered (50% - 74% of items correct)
 ● = Mastered (75% - 99% of items correct)

READING VOCABULARY	WORD MEANING	MULTIMEANING WORDS	AFFIXES	WORDS IN CONTEXT	NUMBER OF STUDENTS: 26			READING COMPREHENSION	STATED INFORMATION	PASSAGE ANALYSIS	CENTRAL IDEA	FORMS, STYLES, TECHNIQUES	CRITICAL THINKING	LANGUAGE ANALYSIS	SENTENCE-LEVEL MECHANICS	PARAGRAPH COHESION/ORGANIZATION	WRITING CONVENTIONS	EDITING SKILLS	LANGUAGE EXPRESSION	USE OF NOUNS, PRONOUNS	USE OF VERBS
					Class	Norm	Diff														
Word Meaning	42%	45%	-3		42	43	-1	42	43	19	19	31	35	35	62	23	38	46	100	95	+5
Multimeaning Words	27	47	-20		19	35	-16	19	32	19	31	35	37	23	25	36	46	77	65	+12	
Affixes	19	26	-7		19	32	-13														
Words in Context	46	27	+19		46	27	+19														
Stated Information					42	43	-1														
Passage Analysis					19	35	-16														
Central Idea					19	32	-13														
Forms, Styles, Techniques					31	31	0														
Critical Thinking					35	37	-2														
Sentence-level Mechanics					62	41	+21														
Paragraph Cohesion/Organization					23	25	-2														
Writing Conventions					38	36	+2														
Editing Skills					46	23	+23														
Use of Nouns, Pronouns					100	95	+5														
Use of Verbs					77	65	+12														

DISRPTO WRD/101291

Questions Pertaining to Score Report Sample B: Class Objective Mastery Report

Please consider the statements below in relation to Score Report B on the facing page. Circle the letter in the second column that corresponds to the degree to which you agree with each statement as applied to Score Report B. (SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree.) If you have no opinion about a statement, circle *n/o*.

<i>Statement</i>	<i>Degree of Agreement</i>					<i>No Opinion</i>	
<i>Score Report B would be useful to a classroom teacher who wants to:</i>							
1. compare his or her students' achievement levels to those of students nationwide.	1.	SD	D	N	A	SA	<i>n/o</i>
2. understand the instructional needs of particular students.	2.	SD	D	N	A	SA	<i>n/o</i>
3. form temporary work groups to focus on individual skill development.	3.	SD	D	N	A	SA	<i>n/o</i>
4. distinguish skill areas that need emphasis from those that do not.	4.	SD	D	N	A	SA	<i>n/o</i>
5. evaluate his or her own teaching effectiveness.	5.	SD	D	N	A	SA	<i>n/o</i>
6. target curriculum areas in which resources and/or teaching methods should be reevaluated.	6.	SD	D	N	A	SA	<i>n/o</i>
7. provide feedback to parents on the skills of their children.	7.	SD	D	N	A	SA	<i>n/o</i>
8. evaluate the effectiveness of the reading/language arts curriculum.	8.	SD	D	N	A	SA	<i>n/o</i>
9. plan instruction that is tailored to individual students' needs.	9.	SD	D	N	A	SA	<i>n/o</i>
<i>On the basis of Score Report B, it is justifiable to conclude that:</i>							
10. compared with students nationwide, this class is above average in "Language Expression" skills.	10.	SD	D	N	A	SA	<i>n/o</i>
11. as a whole, this class has more knowledge gaps on skills covered under "Reading Comprehension" than on those covered under "Reading Vocabulary."	11.	SD	D	N	A	SA	<i>n/o</i>
12. as a whole, this class needs about the same amount of work on "Passage Analysis" and "Central Idea."	12.	SD	D	N	A	SA	<i>n/o</i>
13. as a whole, this class knows more about skills covered under "Words in Context" than skills covered under "Stated Information."	13.	SD	D	N	A	SA	<i>n/o</i>
14. the students in this class spend too much time learning grammar and not enough time actually writing.	14.	SD	D	N	A	SA	<i>n/o</i>
15. Compared to Luisa Ali, George Benne is stronger in "Reading Comprehension" but weaker in "Language Analysis."	15.	SD	D	N	A	SA	<i>n/o</i>

Section III: Background Questions
(Please circle one response letter unless otherwise specified.)

1. To which grade levels are you currently assigned as a teacher? (Circle response letters for all levels that apply.)
A. Kindergarten D. Grade 3 G. Grade 6 J. Other
B. Grade 1 E. Grade 4 H. Grade 7
C. Grade 2 F. Grade 5 I. Grade 8
2. Which of the following describes your primary work environment?
A. Self-contained classroom (teaching the same group of students more than one subject)
B. Departmentalized (teaching the same subject to different groups of students)
C. Multi-setting/itinerant (teaching at more than one school)
D. Administrative (e.g., instructional coordinator)
E. Other
3. What is the approximate population of the municipality (i.e., town or city) in which your school of primary assignment is located?
A. under 25,000 D. 250,000 to 499,999
B. 25,000 to 99,999 E. over 500,000
C. 100,000 to 249,999
4. How many years of teaching experience do you have, counting this year?
A. fewer than 3 C. 10 to 20
B. 3 to 9 D. more than 20
5. During your *preservice* teacher training, how many courses did you take that addressed testing and measurement issues as either the *sole focus* or a *major focus*?
A. none D. 3
B. 1 E. more than 3
C. 2
6. Since you started to teach, how many *inservice courses* (i.e., for credit, not professional development workshops that carried no credits) have you taken that have addressed testing and measurement issues as either the *sole focus* or a *major focus*?
A. none D. 3
B. 1 E. more than 3
C. 2
7. Since you started to teach, how many *inservice workshops* (i.e., not for credit) have you taken that have addressed testing and measurement issues as either the *sole focus* or a *major focus*?
A. none D. 3
B. 1 E. more than 3
C. 2

8. In general, how useful do you think your preservice and inservice preparation has been for dealing with testing and measurement issues that arise on your job?
- A. not at all useful
 - B. rarely useful
 - C. sometimes useful
 - D. generally useful
 - E. very useful
9. Approximately how often do you have to draw upon your knowledge of testing and measurement issues as part of your job?
- A. never
 - B. rarely (1 to 5 times a year)
 - C. occasionally (6 to 15 times a year)
 - D. often (16 to 30 times a year)
 - E. very often (more than 30 times a year)
10. Are you Female or Male?
- A. Female
 - B. Male
11. What is your racial/ethnic background?
- A. American Indian/Alaskan Native
 - B. Asian/Pacific Islander
 - C. Black, Non-Hispanic
 - D. Hispanic
 - E. White, Non-Hispanic
12. With which of the following standardized achievement test batteries (SATBs) are you familiar? (Circle response letters for all that apply.)
- A. California Achievement Test (CAT)
 - B. Comprehensive Test of Basic Skills (CTBS)
 - C. Iowa Test of Basic Skills (ITBS)
 - D. Metropolitan Achievement Test (MAT)
 - E. Stanford Achievement Test (SAT)
13. Which of the following standardized achievement test batteries (SATBs) does your school currently use? (Circle response letters for all that apply.)
- A. California Achievement Test (CAT)
 - B. Comprehensive Test of Basic Skills (CTBS)
 - C. Iowa Test of Basic Skills (ITBS)
 - D. Metropolitan Achievement Test (MAT)
 - E. Stanford Achievement Test (SAT)

THANK YOU AGAIN FOR TAKING THE TIME TO RESPOND TO THIS QUESTIONNAIRE. FEEL FREE TO USE THE BACK COVER FOR ANY ADDITIONAL COMMENTS YOU WOULD LIKE TO MAKE.

APPENDIX B
QUESTIONNAIRE FORM S

**Teacher Questionnaire on the Uses of
Score Reports from Standardized Achievement Test Batteries
(Form S)**

This questionnaire asks teachers for their opinions about the score reports that summarize their students' performance on standardized achievement test batteries. The test batteries of particular interest in this study are the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Test of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), and the Stanford Achievement Test (SAT).

In light of the fact that nearly all school districts administer standardized achievement tests, it is surprising to note that relatively little is known about *teacher* opinions regarding the score reports from these tests. Do teachers find the information in score reports useful? If so, for what purposes? Is the information in these reports understandable? Are the formats of these reports clear and helpful? What changes would teachers like to see? These are the sorts of questions that this study seeks to answer by directly addressing teachers.

You are being asked to respond to this questionnaire because, as an experienced teacher in grades K - 8, you have almost certainly worked with score reports from such tests and formed some opinions about them. Because your opinions are likely to be both informed and practical, they should be of considerable value to those who design and publish tests for classroom use.

We first ask your opinions regarding the potential purposes, in general, of standardized test score reports. Then you are asked to consider two fictitious examples of a widely used type of score report and to answer opinion and interpretive questions about those examples. Finally, we ask some background questions to help describe the sample of teachers who responded to the survey.

In all cases, you are asked to respond on the basis of your own experiences with standardized achievement test score reports and to share your personal opinions and preferences regarding them. Your responses to this questionnaire will be strictly confidential. All information that could identify individual respondents will be eliminated after the questionnaires are returned.

Completion of this questionnaire should take about 30 minutes. We realize that this time expenditure is not trivial in the busy schedule of a teacher, but we hope the knowledge gained on behalf of the education of children will be worth the sacrifice.

If you would like a copy of an executive summary of our findings, please fill out your name and address on the enclosed form. (These forms will be separated from your responses before your responses are tabulated.)

Thank you for your participation.

Ronald K. Hambleton
Professor of Education and Psychology

Edward J. Murphy
Research Associate

University of Massachusetts

Section I: Purposes of Score Reports from Standardized Achievement Test Batteries

Please read the following statements about potential purposes that standardized achievement test battery (SATB) score reports have been said to serve. For each statement, indicate the degree to which you agree that the statement reflects a *valid* purpose that SATB score reports actually serve *in your own experience*. The potential degrees of agreement are as follows:

- SD Strongly Disagree (i.e., the stated purpose is *emphatically not valid* in your experience)
- D Disagree (i.e., the stated purpose is *generally not valid* in your experience)
- N Neutral (i.e., the stated purpose *may or may not be valid* in your experience)
- A Agree (i.e., the stated purpose is *generally valid* in your experience)
- SA Strongly Agree (i.e., the stated purpose is *emphatically valid* in your experience)

If you have no opinion about a statement, circle *n/o*.

Statement (SATB score reports provide useful information for...)	Degree of Agreement (SD = Strongly Disagree to SA = Strongly Agree)	No Opinion
1. helping schools make decisions about placement of individual students into <i>permanent</i> instructional groups (e.g., homogeneous ability groups).	1. SD D N A SA	<i>n/o</i>
2. helping teachers make decisions about placement of individual students into <i>temporary</i> instructional groups (e.g., cooperative learning groups, groups for enrichment or remedial work).	2. SD D N A SA	<i>n/o</i>
3. helping teachers diagnose individual students' strengths, weaknesses, and needs.	3. SD D N A SA	<i>n/o</i>
4. helping teachers keep the pace and level of instruction "on track" with national expectations.	4. SD D N A SA	<i>n/o</i>
5. enabling teachers to measure individual students' growth in particular skills.	5. SD D N A SA	<i>n/o</i>
6. enabling teachers to measure individual students' growth in overall subject areas (e.g., math, language arts).	6. SD D N A SA	<i>n/o</i>
7. enabling teachers to measure group achievement in particular skills over time.	7. SD D N A SA	<i>n/o</i>
8. enabling teachers to measure group achievement in overall subject areas over time.	8. SD D N A SA	<i>n/o</i>
9. enabling teachers to plan instruction that is tailored or adapted to individual students' needs.	9. SD D N A SA	<i>n/o</i>
10. helping teachers establish individual students' grades in class.	10. SD D N A SA	<i>n/o</i>

<i>Statement</i> (SATB score reports provide useful information for...)	<i>Degree of Agreement</i> (SD = Strongly Disagree to SA = Strongly Agree)					<i>No Opinion</i>	
11. enabling schools to make promotion/retention decisions for individual students.	11.	SD	D	N	A	SA	n/o
12. helping teachers and schools evaluate the effectiveness of the curriculum or of curriculum materials.	12.	SD	D	N	A	SA	n/o
13. helping teachers evaluate the effectiveness of their own instructional approaches and strategies.	13.	SD	D	N	A	SA	n/o
14. helping students gain personal insight into their strengths and weaknesses.	14.	SD	D	N	A	SA	n/o
15. helping teachers gain unexpected insights into particular students' hidden talents, achievements, or interests.	15.	SD	D	N	A	SA	n/o
16. enabling teachers to compare individual students' aptitude and achievement levels.	16.	SD	D	N	A	SA	n/o
17. helping teachers eliminate potential sources of personal bias in evaluating their students' abilities by providing an objective form of information on student achievement.	17.	SD	D	N	A	SA	n/o
18. helping teachers explain individual student achievements and needs to parents.	18.	SD	D	N	A	SA	n/o
19. enabling administrators to compare varying programs or approaches being implemented in different classrooms or schools.	19.	SD	D	N	A	SA	n/o
20. helping administrators evaluate the performance of individual teachers.	20.	SD	D	N	A	SA	n/o

If you have any comments, please write them here:

Section II: Sample Score Reports

In this section of the questionnaire you are asked to consider and respond to questions about two sample score reports. These score reports are fictitious; that is, they have been created by the researchers rather than being reproduced exactly from any published test. They are intended, however, to be similar to actual score reports and to present information and format features that are of the sort offered by standardized achievement test publishers.

The two samples are named as follows:

Sample S: Individual Student Objective Report
Sample N: Individual Student Interpretive Report.

The two reports provide imaginary test score information for one student in the fifth month of fifth grade at the fictitious Tyler School. You may assume that the student took a standardized achievement test battery consisting of several subject area subtests, including reading/language arts and mathematics.

Sample S presents information on both the mathematics and reading/language arts subtests for Edmund J. Canton, a student in Mr. or Ms. Hoover's class; Sample N relates to the same two subtests for the same student, but the information is presented in a different format.

Please respond to the questions in this section based on your own personal experience as a teacher.

HSAT Hypothetical Skills Achievement Test

Sample S: Individual Student Objective Report

Individual Student Report			
Student:	Canton	Edmund	J
NS = National Stanine	GE = Grade Equivalent	Class: Hoover	Grade: 5.5
NCE = Normal Curve Equivalent	SS = Scaled Score	School: Tyler	District: Garfield
LP = Local Percentile	NP = National Percentile	City: Anytown	State: Anystate

	NS	GE	NCE	SS	LP	NP	RANGE	NATIONAL PERCENTILE											
								1	2	5	10	20	30	40	50	60	70	80	90
Reading Vocabulary	6	6.1	56	713	70	61	52 - 68	[Graphical representation]											
Reading Comprehension	5	5.6	55	641	61	57	52 - 65	[Graphical representation]											
Language Analysis	5	4.8	47	662	55	53	48 - 68	[Graphical representation]											
Language Expression	5	5.5	55	709	68	56	48 - 62	[Graphical representation]											
Total Reading	5	5.6	55	685	64	58	52 - 65	[Graphical representation]											
Math Concepts	3	4.1	39	678	40	21	18 - 30	[Graphical representation]											
Math Computation	4	4.8	44	707	52	37	30 - 44	[Graphical representation]											
Math Problem Solving	3	3.2	27	649	18	14	10 - 21	[Graphical representation]											
Math Expression	4	4.6	42	700	49	34	30 - 44	[Graphical representation]											
Total Math	4	4.4	40	688	39	28	21 - 32	[Graphical representation]											

INDIVIDUAL OBJECTIVE PERFORMANCE (Percent of Test Questions Answered Correctly)									
Objective	Score	Not Mastered		Partly Mastered		Mastered		Score	
		0	25	50	75	100	0		25
Reading Vocabulary	72	[Graphical]		[Graphical]		[Graphical]		30	
Word Meaning	69	[Graphical]		[Graphical]		[Graphical]		34	
Multimeaning Words	66	[Graphical]		[Graphical]		[Graphical]		25	
Affixes	68	[Graphical]		[Graphical]		[Graphical]		22	
Words in Context	77	[Graphical]		[Graphical]		[Graphical]		30	
Reading Comprehension	66	[Graphical]		[Graphical]		[Graphical]		44	
Stated Information	77	[Graphical]		[Graphical]		[Graphical]		51	
Passage Analysis	62	[Graphical]		[Graphical]		[Graphical]		48	
Central Idea	73	[Graphical]		[Graphical]		[Graphical]		33	
Forms, Styles, Techniques	65	[Graphical]		[Graphical]		[Graphical]		40	
Critical Thinking	61	[Graphical]		[Graphical]		[Graphical]		36	
Language Analysis	72	[Graphical]		[Graphical]		[Graphical]		24	
Sentence-level Mechanics	78	[Graphical]		[Graphical]		[Graphical]		28	
Paragraph Cohesion/Organization	68	[Graphical]		[Graphical]		[Graphical]		26	
Writing Conventions	76	[Graphical]		[Graphical]		[Graphical]		20	
Editing Skills	72	[Graphical]		[Graphical]		[Graphical]		18	
Language Expression	74	[Graphical]		[Graphical]		[Graphical]		42	
Use of Nouns, Pronouns	78	[Graphical]		[Graphical]		[Graphical]		45	
Use of Verbs	73	[Graphical]		[Graphical]		[Graphical]		40	

Questions Pertaining to Score Report Sample S: Individual Student Objective Report

Please consider the statements below in relation to Score Report S on the facing page. Circle the letter in the second column that corresponds to the degree to which you agree with each statement as applied to Score Report S. (SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree.) If you have no opinion about a statement, circle *n/o*.

<i>Statement</i>	<i>Degree of Agreement</i>	<i>No Opinion</i>
<i>Score Report S would be useful to a classroom teacher who wants to:</i>		
1. compare this student's achievement levels to those of students nationwide.	1. SD D N A SA	<i>n/o</i>
2. understand this student's academic strengths and weaknesses.	2. SD D N A SA	<i>n/o</i>
3. create an instructional plan targeted to this student's needs.	3. SD D N A SA	<i>n/o</i>
4. find out the difficulty level of the reading materials with which this student will be comfortable.	4. SD D N A SA	<i>n/o</i>
5. learn the grade levels at which this student is performing in the skill areas covered by the test.	5. SD D N A SA	<i>n/o</i>
6. help this student understand his own test performance.	6. SD D N A SA	<i>n/o</i>
7. discuss this student's test performance with his parents.	7. SD D N A SA	<i>n/o</i>
8. evaluate the effectiveness of instructional strategies, curriculum, and/or resources now in use in this classroom.	8. SD D N A SA	<i>n/o</i>
9. set up groups of students to work together on specific skills.	9. SD D N A SA	<i>n/o</i>
<i>On the basis of Score Report S, it is justifiable to conclude that:</i>		
10. this student has the reading vocabulary of a beginning sixth grader.	10. SD D N A SA	<i>n/o</i>
11. on the skills covered under "Total Reading," this student performed better than 58 percent of the nation's fifth graders.	11. SD D N A SA	<i>n/o</i>
12. compared with the nation's fifth graders, this student is in the lowest quartile on the skills covered under "Math Concepts."	12. SD D N A SA	<i>n/o</i>
13. this student knows more about the skills covered under "Sentence-level Mechanics" than those covered under "Writing Conventions."	13. SD D N A SA	<i>n/o</i>
14. this student knows more about the skills covered under "Math Computation" than those covered under "Math Expression."	14. SD D N A SA	<i>n/o</i>
15. this student is in the "Mastered" category on the skills covered under "Use of Nouns, Pronouns."	15. SD D N A SA	<i>n/o</i>

HSAT Hypothetical Skills Achievement Test

Sample N: Individual Student Interpretive Report

Individual Student Report		
Student:	Canton Edmund J	
	NS = National Stanine	GE = Grade Equivalent
	NCE = Normal Curve Equivalent	SS = Scaled Score
	LP = Local Percentile	NP = National Percentile
		Class: Hoover
		School: Tyler
		City: Anytown
		District: Garfield
		State: Anystate
		Grade: 5.5

	NS	GE	NCE	SS	LP	NP	RANGE	NATIONAL PERCENTILE																		
								1	2	5	10	20	30	40	50	60	70	80	90	95	98	99				
Reading Vocabulary	6	6.1	56	713	70	61	52 - 68																			
Reading Comprehension	5	5.6	55	641	61	57	52 - 65																			
Language Analysis	5	4.8	47	662	55	53	48 - 68																			
Language Expression	5	5.5	55	709	68	56	48 - 62																			
Total Reading	5	5.6	55	685	64	58	52 - 65																			
Math Concepts	3	4.1	39	678	40	21	18 - 30																			
Math Computation	4	4.8	44	707	52	37	30 - 44																			
Math Problem Solving	3	3.2	27	649	18	14	10 - 21																			
Math Expression	4	4.6	42	700	49	34	30 - 44																			
Total Math	4	4.4	40	688	39	28	21 - 32																			

Detailed Interpretation of Scores

National Norms

This student's test performance can be compared with that of the national norm group by referring to the National Percentile (NP) column above. This student performed above the national average (the 50th percentile) in all reading subtests and in Total Reading. In Total Reading, the student's achievement was better than approximately 58 percent of the nation's fifth graders. The student performed below the national average (the 50th percentile) in all mathematics subtests and in Total Math. In Total Math, the student's achievement was better than approximately 28 percent of the nation's fifth graders.

Performance on Tested Objectives

The Reading portion of the test covered four areas: Reading Vocabulary (RV), Reading Comprehension (RC), Language Analysis (LA), and Language Expression (LE). In the reading portion, the student achieved scores indicating at least partial mastery of all objectives. The student appears to be strongest in Words in Context (RV), Stated Information (RC), Sentence-level Mechanics (LA), and Use of Nouns and Pronouns (LE). The student may need further instruction to improve skills in Passage Analysis (RC) and Critical Thinking (RC).

The Mathematics portion of the test covered four areas: Math Concepts (MC), Math Computation (MCom), Math Problem Solving (MPS), and Mathematical Expression (ME). In the math portion, the student achieved scores indicating partial mastery of one objective: Add/Subtract Whole Numbers. All other scores appear to indicate non-mastery of objective content. The student appears to have performed more strongly in the general areas of Mathematical Computation and Mathematical Expression than in Mathematical Concepts and Mathematical Problem Solving, but the student may need further instruction in all areas of math covered by the test. Areas of particular need appear to be Critical Thinking Problems (MPS), Multistep Problems (MPS), and Decimals and Percent (MC).

Questions Pertaining to Score Report Sample N: Individual Student Interpretive Report

Please consider the statements below in relation to Score Report N on the facing page. Circle the letter in the second column that corresponds to the degree to which you agree with each statement as applied to Score Report N. (SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree.) If you have no opinion about a statement, circle *n/o*.

<i>Statement</i>	<i>Degree of Agreement</i>	<i>No Opinion</i>
<i>Score Report N would be useful to a classroom teacher who wants to:</i>		
1. compare this student's achievement levels to those of students nationwide.	1. SD D N A SA	<i>n/o</i>
2. understand this student's academic strengths and weaknesses.	2. SD D N A SA	<i>n/o</i>
3. create an instructional plan targeted to this student's needs.	3. SD D N A SA	<i>n/o</i>
4. find out the difficulty level of the reading materials with which this student will be comfortable.	4. SD D N A SA	<i>n/o</i>
5. learn the grade levels at which this student is performing in the skill areas covered by the test.	5. SD D N A SA	<i>n/o</i>
6. help this student understand his own test performance.	6. SD D N A SA	<i>n/o</i>
7. discuss this student's test performance with his parents.	7. SD D N A SA	<i>n/o</i>
8. evaluate the effectiveness of instructional strategies, curriculum, and/or resources now in use in this classroom.	8. SD D N A SA	<i>n/o</i>
9. set up groups of students to work together on specific skills.	9. SD D N A SA	<i>n/o</i>
<i>On the basis of Score Report N, it is justifiable to conclude that:</i>		
10. this student has the math problem solving skills of a third grader.	10. SD D N A SA	<i>n/o</i>
11. compared with the nation's fifth graders, this student is above average on the skills covered under "Language Analysis."	11. SD D N A SA	<i>n/o</i>
12. on the skills covered under "Total Math," this student performed better than 28 percent of the nation's fifth graders.	12. SD D N A SA	<i>n/o</i>
13. this student knows more about the skills covered under "Language Expression" than those covered under "Language Analysis."	13. SD D N A SA	<i>n/o</i>
14. this student knows more about the skills covered under "Add/Subtract Whole Numbers" than those covered under "Use of Math Symbols/Terms."	14. SD D N A SA	<i>n/o</i>
15. this student is in the "Not Mastered" category on the skills covered under "Multiply/Divide Whole Numbers."	15. SD D N A SA	<i>n/o</i>

Section III: Background Questions
(Please circle one response letter unless otherwise specified.)

1. To which grade levels are you currently assigned as a teacher? (Circle response letters for all levels that apply.)
A. Kindergarten D. Grade 3 G. Grade 6 J. Other
B. Grade 1 E. Grade 4 H. Grade 7
C. Grade 2 F. Grade 5 I. Grade 8

2. Which of the following describes your primary work environment?
A. Self-contained classroom (teaching the same group of students more than one subject)
B. Departmentalized (teaching the same subject to different groups of students)
C. Multi-setting/itinerant (teaching at more than one school)
D. Administrative (e.g., instructional coordinator)
E. Other

3. What is the approximate population of the municipality (i.e., town or city) in which your school of primary assignment is located?
A. under 25,000 D. 250,000 to 499,999
B. 25,000 to 99,999 E. over 500,000
C. 100,000 to 249,999

4. How many years of teaching experience do you have, counting this year?
A. fewer than 3 C. 10 to 20
B. 3 to 9 D. more than 20

5. During your *preservice* teacher training, how many courses did you take that addressed testing and measurement issues as either the *sole focus* or a *major focus*?
A. none D. 3
B. 1 E. more than 3
C. 2

6. Since you started to teach, how many *inservice courses* (i.e., for credit, not professional development workshops that carried no credits) have you taken that have addressed testing and measurement issues as either the *sole focus* or a *major focus*?
A. none D. 3
B. 1 E. more than 3
C. 2

7. Since you started to teach, how many *inservice workshops* (i.e., not for credit) have you taken that have addressed testing and measurement issues as either the *sole focus* or a *major focus*?
A. none D. 3
B. 1 E. more than 3
C. 2

8. In general, how useful do you think your preservice and inservice preparation has been for dealing with testing and measurement issues that arise on your job?
- A. not at all useful
 - B. rarely useful
 - C. sometimes useful
 - D. generally useful
 - E. very useful
9. Approximately how often do you have to draw upon your knowledge of testing and measurement issues as part of your job?
- A. never
 - B. rarely (1 to 5 times a year)
 - C. occasionally (6 to 15 times a year)
 - D. often (16 to 30 times a year)
 - E. very often (more than 30 times a year)
10. Are you Female or Male?
- A. Female
 - B. Male
11. What is your racial/ethnic background?
- A. American Indian/Alaskan Native
 - B. Asian/Pacific Islander
 - C. Black, Non-Hispanic
 - D. Hispanic
 - E. White, Non-Hispanic
12. With which of the following standardized achievement test batteries (SATBs) are you familiar? (Circle response letters for all that apply.)
- A. California Achievement Test (CAT)
 - B. Comprehensive Test of Basic Skills (CTBS)
 - C. Iowa Test of Basic Skills (ITBS)
 - D. Metropolitan Achievement Test (MAT)
 - E. Stanford Achievement Test (SAT)
13. Which of the following standardized achievement test batteries (SATBs) does your school currently use? (Circle response letters for all that apply.)
- A. California Achievement Test (CAT)
 - B. Comprehensive Test of Basic Skills (CTBS)
 - C. Iowa Test of Basic Skills (ITBS)
 - D. Metropolitan Achievement Test (MAT)
 - E. Stanford Achievement Test (SAT)

THANK YOU AGAIN FOR TAKING THE TIME TO RESPOND TO THIS QUESTIONNAIRE. FEEL FREE TO USE THE BACK COVER FOR ANY ADDITIONAL COMMENTS YOU WOULD LIKE TO MAKE.

APPENDIX C
SUPPORT MATERIALS SENT TO SURVEY LIAISONS

Post Office Box 226
Amherst, MA 01004
November 20, 1991

Houston, TX 77009

Dear Mr.

Thank you for agreeing to serve as a distribution coordinator for the survey instruments enclosed. Accompanying this letter you should find the following materials.

- one "Acknowledgement of Receipt" form, which I ask you to complete and return to me as soon as possible
- one prepaid white business-size envelope for your use in returning the "Acknowledgement of Receipt" form
- one copy of an abstract of the study for your use
- one draft of a cover letter to survey recipients from you, which you may or may not choose to use
- one draft of a follow-up letter to survey recipients from you, which you may or may not choose to use
- _____ unsealed manila envelopes for survey recipients, each containing one cover letter from me, one questionnaire (one of four different forms), a letter from Dr. Nolan Wood of the Texas Education Agency encouraging participation, and a form for participants to use in requesting an executive summary of the study's results
- _____ return Federal Express envelopes and prepaid airbills, for your use in returning completed surveys to me
- one sheet containing a suggested procedure to use in distributing, collecting, and returning the surveys

Please check that you have received these materials and read the suggested procedure. Then return the "Acknowledgement of Receipt" form in the enclosed white business-size envelope.

I very much appreciate your help.

Sincerely,

Edward J. Murphy

ACKNOWLEDGEMENT OF RECEIPT

Dear Mr. Murphy:

_____ I have received the package containing the surveys for your study. All materials on your checklist were delivered.

_____ I need the following materials: _____

**ELEMENTARY SCHOOL TEACHERS' OPINIONS REGARDING
THE PURPOSES AND INTERPRETATION OF SCORE REPORTS
FROM STANDARDIZED ACHIEVEMENT TEST BATTERIES**

Edward J. Murphy

ABSTRACT

Standardized achievement test survey batteries are widely used in elementary schools in the United States. Such tests appear likely to remain prevalent so long as they are regarded as useful by their major audiences. Despite recent criticisms, achievement test batteries are apparently still seen as useful for some purposes by at least some influential users.

Classroom teachers play a critical role in the use of standardized achievement tests. Teachers not only administer such tests, they also are expected to make use of the information that derives from them. To make good use of test results, teachers must first regard the tests as serving useful purposes; then they must understand how to interpret the information presented in the score reports so as to accomplish those purposes. However, because of practical limitations in their teacher education coursework, teachers may not feel adequately prepared to interpret highly technical testing information in score reports.

This state of affairs places great importance on the test score report as the central communication device between the test and the teacher. If teachers are unlikely to have a great deal of technical knowledge to bring to the interpretation of the score report, it can be argued that the score report ought to be designed to communicate clearly and accurately the information the teacher regards as useful.

In this proposal a study is described which attempts to ascertain by means of a questionnaire the purposes for which elementary teachers believe standardized achievement test survey battery score reports are potentially useful, and the content and format of score reports that teachers believe would fulfill those purposes. In addition, teacher interpretations of various test score report contents and formats are examined.

The importance of this study lies in its focus on achievement test score reports as a critical form of communication between test publishers and one important type of test user: the elementary classroom teacher. Teachers' support for achievement batteries and understanding of score reports are essential if test results are to be put to proper use in classroom decision-making. This study aims to increase knowledge of teachers' opinions regarding the proper purposes of score reports and teachers' needs and preferences regarding the content and format of these important messages so that score reports can be designed to communicate their critical messages effectively.

Dear Colleague:

I have agreed to distribute the enclosed questionnaire to you as part of a study of teachers' opinions regarding score reports from standardized achievement tests. This study is serious in intent and, because it concerns an issue of great interest to classroom teachers, I encourage you to spend the time to complete the questionnaire.

The materials enclosed with the survey contain all the information you should need to complete it. If you have questions, I will be happy to try to answer them. I also have a copy of an abstract of the study that describes it in more detail, if you would like to see it.

Please return the completed questionnaire to me in its envelope **WITHIN ABOUT TWO WEEKS**. I will then send it back to the researchers.

Thank you for your time.

Dear Colleague:

About two weeks ago you received from me a copy of a questionnaire seeking teachers' opinions about score reports from standardized achievement tests. As I explained then, I have agreed to collect all completed questionnaires and send them back to the researchers.

I am sending this follow-up note to all recipients of the questionnaire. If you have not returned your questionnaire, please make an effort to complete it and return it to me as soon as you can. If you need another copy of the questionnaire, please call me.

Thank you.

QUESTIONNAIRE DISTRIBUTION COORDINATORS
Suggested Procedure for Distributing, Collecting, and Returning the Surveys

1. Please read the draft cover letter to survey recipients from you ("I have agreed to distribute...") and decide whether or not to use it. If you decide to use it, please sign it, make copies, and place one copy in each manila envelope. If you would prefer to compose a note of your own, feel free to use the draft letter as a base. *Do not seal the manila envelopes unless you intend to mail them individually to survey recipients via the Postal Service.*
2. Distribute one manila envelope to each survey recipient whom you have identified. Eligible survey recipients are elementary-level teachers (defined as K - 8) who may have had experience with the use of score reports from standardized achievement tests. (NOTE: There are four different forms of the questionnaire in this study, so not every recipient will receive the same survey. To ensure randomization, please distribute the survey envelopes in the order in which they were packed.)
3. Please complete a survey yourself if you are eligible.
4. Recipients are instructed to return the completed questionnaires to you within approximately two weeks. When the questionnaires start to come in, transfer them from their manila envelopes to one of the Federal Express envelopes provided or to a box, if that is more convenient. Please wait until the Federal Express envelope/box is full before sending it back to me.
5. To send the Federal Express envelope or box, use one of the airbills enclosed. All you have to fill out is the sender's address; everything else has already been filled out and the bill will be paid by me. Place the completed airbill in the clear plastic sleeve, fasten the sleeve to the envelope or box, and call Federal Express for pickup.
6. *Please discard everything except the questionnaires, respondents' request forms for executive summaries, and any notes the recipients addressed to me; I do not need totally blank questionnaires, manila envelopes, or other accompanying materials back.*
7. About two or three weeks from the day you distributed the envelopes, please send a follow-up reminder to all recipients asking them to return the survey if they have not already done so. You may want to use the wording of the enclosed draft follow-up letter ("About two weeks ago.."). Please wait before sending me the last Federal Express mailing to be sure that all returns are in.

Thank you again for your help. If you need anything else or have any questions, please get in touch with me. You may call collect at the number below between 8:30 A.M. and 5:00 P.M. EST.

Edward J. Murphy
P.O. Box 226
Amherst, MA 01004
(413) 256-0444

APPENDIX D
COVER MATERIALS TO RESPONDENTS

Post Office Box 226
Amherst, MA 01004
(413) 256-0444

Dear Educator:

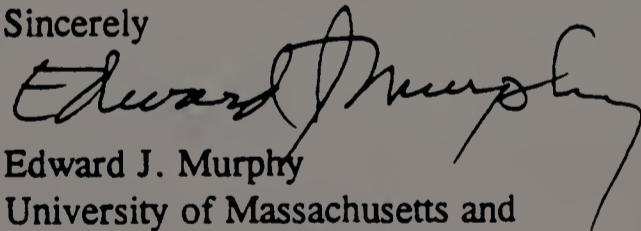
You are being asked to participate in a study regarding an issue of importance to classroom teachers and their students. The study seeks to gather teachers' opinions about the score reports that summarize the results of the standardized achievement tests that students take in the schools.

Enclosed you will find a questionnaire that you are asked to complete. It should take no more than a half-hour to fill out. I know that your time is limited, but I hope you will complete the questionnaire because of its potential value to educators and students in this country. I have little to offer as an inducement beyond my professional gratitude and the courtesy of an executive summary of the results of the study, if you would like one.

When you have completed the questionnaire, please return it to the person from whom you received it. That person has agreed to serve as my contact and will send all completed surveys back to me.

Thank you for your participation.

Sincerely



Edward J. Murphy
University of Massachusetts and
National Evaluation Systems



Texas Education Agency

1701 NORTH CONGRESS AVENUE AUSTIN, TEXAS 78701-1494 (512) 463-9734

TO THE EDUCATOR ADDRESSED:

You are invited to participate in a dissertation study being conducted by Edward Murphy of National Evaluation Systems, Inc. (NES). The Texas Education Agency (TEA) has had a working relationship with NES since 1984 and has worked extensively with Mr. Murphy over the years. We believe his research will contribute to our knowledge about assessment and student achievement.

We hope that you will consider being involved in this project. Participation should take only a half hour of any teacher's time, but the benefits of this study could extend well beyond that commitment of time.

Please refer to information provided by Mr. Murphy for details concerning his study. Thank you for considering becoming involved.

Sincerely yours,

A handwritten signature in cursive script that reads "Nolan Wood".

Nolan Wood, Director
Division of Teacher Assessment

REQUEST FOR EXECUTIVE SUMMARY

Please send me a copy of the executive summary of the results of your study.

Name: _____

Address: _____

(Return this form with your completed questionnaire to the person from whom you received the survey.)

APPENDIX E
LETTER TO MASSACHUSETTS RESPONDENTS



UNIVERSITY OF MASSACHUSETTS
AT AMHERST

Office of the Dean
School of Education

Room 124 Furber Hall
Amherst, MA 01003
(413) 545-3233

December 3, 1991

Dear Teacher:

I am writing on behalf of Ronald K. Hambleton and Edward J. Murphy of our Research and Evaluation Methods Program in the School of Education. They are conducting a study concerning the effectiveness of reporting methods used with standardized achievement tests. The study is focused on the opinions of teachers about the current report forms: How useful are the reports? How clear is the information in the reports?

I enthusiastically support this study because it promises to shed light on an aspect of standardized testing that has received little research attention -- the score report. And yet, if the many millions of standardized tests that are administered in our schools each year are to serve useful purposes, the score report must communicate clearly and succinctly with teachers. The approach this study takes is to ask teachers for their considered opinions about the ways in which score reports communicate their messages.

Elementary-level teachers (K through 8) in Massachusetts, Illinois, and Texas are being asked to participate in this study. Only with your cooperation can the needed information be obtained. I ask you to take the time from your busy schedule to complete the enclosed survey.

The information enclosed with the survey describes the procedure for completing and returning it. Your responses will be entirely confidential, and, of course, your participation is voluntary.

The results, once analyzed, will add to the knowledge of test developers and others interested in the uses of standardized achievement tests. An executive summary of results is offered to any participant who would like one; simply fill out the enclosed "Request for Executive Summary" and return it with your completed survey to the person from whom you received the survey packet.

Thank you for your assistance.

Sincerely,

Bailey W. Jackson
Dean

BIBLIOGRAPHY

- Airasian, P. W. (1980). Uses and misuses of standardized tests. *Measurement in Education, 10*, 1-8.
- Airasian, P. W. (1991). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice, 10*, 13-16.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1985). *Standards for educational and psychological testing*. Washington, D. C.: American Psychological Association.
- Anastasi, A. (1985). Interpreting results from multiscore batteries. *Journal of Counseling and Development, 64*, 84-6.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan Publishing Co., Inc.
- Anderson, B. (1982). Test use today in elementary and secondary schools. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies. Part II*. Washington, D. C.: National Academy Press.
- Anderson, S. B. (1981). Parents and standardized tests. *Peabody Journal of Education, 58*, 96-107.
- Anttonen, R. G., & Fleming, E. S. (1976). Standardized test information: Does it make a difference in black student performance? *Journal of Educational Research, 70*, 26-31.
- Betts, G. L. (1950). Suggestions for a better interpretation and use of standardized achievement tests. *Education, 71*, 217-21.
- Birnebaum, M., & Shaw, D. J. (1985). Task specification chart: A key to better understanding of test results. *Journal of Educational Measurement, 22*, 219-30.
- Boegli, R. G., Whately, W. W., & Ward, P. E. (1977). Standardized test data for diagnostic-prescriptive instruction. *School Counselor, 24*, 270-2.
- Bohning, G. (1979a). A profile for communicating achievement test results to children. *Elementary School Guidance and Counseling, 13*, 256-60.
- Bohning, G. (1979b). A profile graph for interpreting the Detroit tests of learning aptitude. *Psychology in the Schools, 16*, 338-41.
- Borg, W. R. & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). New York: Longman
- Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: Holt, Rinehart, and Winston.

- Bruno, J. E., Holland, J. R., & Ward, J. W. (1988). Enhancing academic support services for special action students: An application of information referenced testing. *Measurement and Evaluation in Counseling and Development, 21*, 5-15.
- Bruno, J. E. (1989). Monitoring the academic progress of low-achieving students: An analysis of right-wrong (R-W) versus information referenced (MCW-APM) formative and summative evaluation procedures. *Journal of Research and Development in Education, 23*, 51-61.
- Buros, O. K. (1979). Fifty years in testing: Some reminiscences, criticism, and suggestions. *Educational Researcher, 6*, 9-15.
- Campbell, H. H. (1946). Making test results function in teaching. *Education, 66*, 411-5.
- Campbell, N. J. (1981). Interpreting scores from standardized tests. *Clearing House, 55*, 155-7.
- Clark, P. I. (1957). Teachers' use and understanding of tests. *Journal of Education, 139*, 23-38.
- Conant, J. B. (1963). *The education of American teachers*. New York: McGraw-Hill.
- Cox, R. C., & Sterrett, B. G. (1970). A model for increasing the meaning of standardized test scores. *Journal of Educational Measurement, 7*, 227-8.
- Crook, F. E. (1959). Elementary school testing program problems and practices. *Teachers College Record, 61*, 76-85.
- Culyer, R. C. (1982). Interpreting achievement test data: Some areas of concern. *Clearing House, 55*, 374-80.
- Cummings, O. W. (1981). Student-centered test interpretation: An active technique. *School Counselor, 28*, 267-72.
- Cummings, O. W., & Stinard, T. A. (1983). The impact of student-centered test interpretation on teachers and students. *School Counselor, 31*, 66-74.
- Cunningham, W. (1968). How to explain test scores to parents. *School Management, 12*, 86-7
- Curtis, M. E., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. *Journal of Educational Measurement, 20*, 133-47.
- Daggett, C. J. (1934). How teachers go astray. *Educational Administration and Supervision, 20*, 209-15.
- Davis, F. B. (1962). Testing and the use of test results: Test-score interpretation. *Review of Educational Research, 32*, 5-14.

- Dreher, M. J., & Singer, H. (1985). Parents' attitudes toward reports of standardized test results. *Reading Teacher*, 38, 624-32.
- Durost, W. N. (1959). Problems in in-service training of teachers in the use of measurement and evaluation techniques. *Sixteenth Yearbook of the National Council on Measurements Used in Education*, 31-33.
- Ebel, R. L. (1961). Improving the competence of teachers in educational measurement. *Clearing House*, 36, 67-71.
- Ebel, R. L. (1976). A paradox of educational testing. *Measurement in Education*, 7, 4.
- Ebel, R. L., & Hill, R. E. (1959). Development and applications of tests of educational achievement. *Review of Educational Research*, 29, 42-56.
- Fleming, M. (1971). Standardized tests revisited. *School Counselor*, 19, 71-2.
- Flood, J., & Lapp, D. (1989). Reporting reading progress: A comparison portfolio for parents. *Reading Teacher*, 42, 508-14.
- Fredrickson, R. H., & Marchie, H. E. (1966). Teachers view test results. *Clearing House*, 40, 357-8.
- Furlong, F., & Miller, W. (1978). DIAGNOSE: Computer-based reporting of criterion-referenced test results. *Educational Technology*, 8, 37-9.
- Gardner, E. F. (1977). Interpreting achievement profiles: Uses and warnings. *Journal of Research and Development in Education*, 10, 51-63.
- Gardner, E. F. (1978). Bias. *Measurement in Education*, 9, 3.
- Gardner, E. (1982). Some aspects of the use and misuse of standardized aptitude and achievement tests. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies. Part II*. Washington, D. C.: National Academy Press.
- Genck, F. H. (1989). To see how well your schools are doing, scrutinize test scores. *American School Board Journal*, 176, 27-8.
- Goldman, L. (1972). Tests and counseling: The marriage that failed. *Measurement and Evaluation in Guidance*, 4, 213-20.
- Goslin, D. A. (1967). *Teachers and testing*. New York: Russell Sage.
- Green, D. R. (1987). A guide for interpreting standardized test scores. *NASSP Bulletin*, 71, 23-4.
- Gullickson, A. R. (1982). The practice of testing in elementary and secondary schools. Paper presented at the 1982 Rural Education Conference, Kansas State University, Manhattan, KS. (ERIC Document Reproduction Service No. 229 391).

- Gullickson, A. R. (1984). Teacher perspectives on their instructional use of tests. *Journal of Educational Research*, 77, 244-8.
- Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement*, 23, 347-54.
- Hagen, E., & Lindberg, L. (1963). Staff competence in testing. In N. B. Henry & H. G. Richey (Eds.), *The impact and improvement of school testing programs: 62nd yearbook of the National Society for the Study of Education*. Chicago, IL: University of Chicago Press.
- Hall, E. C. (1954). Proper use of test results. *Elementary School Journal*, 54, 450-5.
- Hambleton, R. K. (1980). Latent ability scales: Interpretations and uses. In S. T. Mayo (Ed.), *Interpreting test performance*. San Francisco, CA: Jossey-Bass.
- Hanna, G. S. (1988). Using percentile bands for meaningful descriptive test score interpretations. *Journal of Counseling Development*, 66, 477-83.
- Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement*, 20, 191-206.
- Hills, J. R. (1991). Apathy concerning grading and testing. *Phi Delta Kappan*, 72, 540-5.
- Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement*, 3, 8-18.
- Hoover, H. D., & Fleetwood, G. R. (1977). A model for increasing decision making potential through standardized test scores. *Psychology in the Schools*, 14, 67-71.
- Hostrop, G. J. (1966). Achievement tests: A reform in reporting needed. *Phi Delta Kappan*, 47, 557.
- Huba, G. J. (1986). Interval banded profile analysis: A method for matching score profiles to "soft" prototypic patterns. *Educational and Psychological Measurement*, 46, 565-70.
- Huebner, E. S. (1987). Teachers' special education decisions: Does test information make a difference? *Journal of Educational Research*, 80, 202-5.
- Huebner, E. S. (1988). Bias in teachers' special education decisions as a function of test score reporting format. *Journal of Educational Research*, 81, 217-20.
- Huebner, E. S. (1989). Errors in decision-making: A comparison of school psychologists' interpretations of grade equivalents, percentiles, and deviation IQs. *School Psychology Review*, 18, 51-5.
- Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice*, 10, 16-18.

- Jackson, C. D. (1975). On the report of the Ad Hoc Committee on Educational Uses of Tests with Disadvantaged Students: Another psychological view from the Association of Black Psychologists. *American Psychologist*, 30, 86-90.
- Jacobs, H. L. (1988). Of birchrods and percentiles. *Contemporary Education*, 59, 162-4.
- Jones, G., & Galbraith, A. (1941). Interpretation of standardized tests. *School and Society*, 54, 224-7.
- Killian, C. R. (1983). Standardized testing and computer technology: New opportunities for improvement. *Educational Technology*, 23, 30-1.
- Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, 25, 97-110.
- Kolstol, R. H. (1967). Uses of test results. *Childhood Education*, 44, 165-7.
- Ladd, E. M. (1971). More than scores from tests. *Reading Teacher*, 24, 305-11.
- Lapointe, A. (1987). Test results provide data useful to educators planning to improve schools. *NASSP Bulletin*, 71, 73-8.
- Leiter, K. (1976). Teachers' use of background knowledge to interpret test scores. *Sociology of Education*, 49, 59-65,
- Lenke, J. M., & Beck, M. D. (1980). The ways and means of score interpretation. In S. T. Mayo (Ed.), *Interpreting test performance*. San Francisco, CA: Jossey-Bass.
- LeSage, W. (1973). Student testing profiles keep the data together. *Instructor*, 82, 49-50.
- Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement*, 20, 179-89.
- Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. *Teachers College Record*, 91, 422-36.
- Linn, R. L., & Hambleton, R. K. (1991). Customized tests and customized norms. *Applied Measurement in Education*, 4, 185-207.
- Lyman, H. B. (1974). Know the score before the game begins: Tests in counseling. *Measurement and Evaluation in Guidance*, 7, 150-6.
- Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66, 611-17.
- Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65, 29-46.
- Madsen, I. N. (1929). Participation in testing programs by the classroom teacher. *Educational Administration and Supervision*, 15, 117-26.

- Marston, P. (1991). School test for 7-year olds to be changed. In the (London) *Daily Telegraph*, May 8, 1991, p. 4. London, England: The Daily Telegraph Co.
- Mathews, W. M. (1972). Development of computer-generated verbal-format testing reports. *Journal of Educational Data Processing*, 9, 1-11.
- Mathews, W. M. (1973). Narrative format testing reports and traditional testing reports: A comparative study. *Journal of Educational Measurement*, 10, 171-8.
- Maxwell, S. E., & Howard, G. S. (1981). Change scores—necessarily anathema? *Educational and Psychological Measurement*, 41, 747-56.
- Mayo, S. T. (1959). Testing and the use of test results. *Review of Educational Research*, 29, 5-14.
- Mayo, S. T. (1964). What experts think teachers ought to know about educational measurement. *Journal of Educational Measurement*, 1, 79-86.
- Mayo, S. T. (1967). Pre-service preparation of teachers in educational measurement. (Contract No. OE 4-10-011). Chicago, IL: Loyola University.
- McArthur, D.L., & Choppin, B. H. (1984). Computerized diagnostic testing. *Journal of Educational Measurement*, 21, 391-7.
- Mehrens, W. H. (1967). Consequences of misusing test results. *National Elementary Principal*, 47, 62-4.
- Mehrens, W. H., & Lehmann, I. J. (1985). Interpreting test scores to clients: What score should one use? *Journal of Counseling and Development*, 63, 317-20.
- Mehrens, W. H., & Lehmann, I. J. (1987). *Using standardized tests in education* (4th ed.). New York: Longman.
- Morse, P. K. (1964). Reporting test results: Percentile bands vs. percentile ranks. *Journal of Educational Measurement*, 1, 139-42.
- Moughamian, H. (1965). General overview of trends in testing. *Review of Educational Research*, 35, 5-16.
- Neill, D., & Medina, N. (1989). Standardized testing: harmful to educational health. *Phi Delta Kappan*, 70, 688-702.
- Noeth, R. J. (1976). Converting student data to counseling information. *Measurement and Evaluation in Guidance*, 9, 60-9.
- Noll, V. H. (1955). Requirements in educational measurement for prospective teachers. *School and Society*, 82, 88-90.

- O'Sullivan, R. G., & Chalnack, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*, 17-19.
- Peckens, R. G., & Bennett, L. M. (1968). A study of the effectiveness of the secondary school counselor in test interpretation. *School Counselor, 15*, 203-8.
- Phillips, B. N., & Weathers, G. (1958). Analysis of errors made in scoring standardized tests. *Educational and Psychological Measurement, 18*, 563-7.
- Pikulski, J. J. (1990). The role of tests in a literacy assessment program. *Reading Teacher, 43*, 686-8.
- Plumleigh, G. (1977). Achievement test scores: Are they of any use? *Phi Delta Kappan, 58*, 614-15.
- Popham, W. J., & Hambleton, R. K. (1990). Can you pass the test on testing? *Principal, 69*, 38-9.
- Pounds, W. J., & Hawkins, M. L. (1968). How much should you tell parents? *Instructor, 78*, 55-57.
- Prediger, D. J. (1971a). Converting test data to counseling information: A system trial with feedback. *Journal of Educational Measurement, 8*, 161-9.
- Prediger, D. J. (1971b). Data-information conversion in test interpretation. *Journal of Counseling Psychology, 18*, 306-13.
- Prescott, G. A. (1971). Criterion-referenced test interpretation in reading. *Reading Teacher, 24*, 347-54.
- Putt, R. C., & Ray, D. D. (1965). Putting test results to work. *Elementary School Journal, 65*, 439-44.
- Rapp, M. L., & Haggart, S. A. (1973). Idiographic analysis of achievement measures. *Educational Technology, 13*, 23-6.
- Ravitch, D. (1983-84). The uses and misuses of tests. *College Board Review, 130*, 22-6.
- Rayborn, R. R. (1989). Learn what test scores really tell you about your schools. *American School Board Journal, 176*, 34-5.
- Resnick, D. P. (1981). Testing in America: A supportive environment. *Phi Delta Kappan, 62*, 625-8.
- Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies. Part II*. Washington, D. C.: National Academy Press.

- Reynolds, C. R. (1981). The fallacy of "two years below grade level for age" as a diagnostic criterion for reading disorders. *Journal of School Psychology, 19*, 350-8.
- Roberge, J.J., & Kubiniec, C. M. (1970). A computer program for verbal feedback on multiple-choice examinations. *Educational and Psychological Measurement, 30*, 175-81.
- Roeber, E., & Dutcher, P. (1989). Michigan's innovative assessment of reading. *Educational Leadership, 46*, 64-9.
- Rogers, A. C. (1961). *Graphic charts handbook*. Washington, D. C.: Public Affairs Press.
- Roid, G. H. (1984). Computer technology in testing. In B. S. Plake & J. C. Witts (Eds.), *The future of testing: The second Buros-Nebraska symposium on measurement and testing*. Hillsdale, NJ: Erlbaum.
- Ross, R. P. (1990). Consistency among school psychologists in evaluating discrepancy scores: A preliminary study. *Learning Disabilities Quarterly, 13*, 209-19.
- Rost, P. (1973). Useful interpretation of standardized tests. *Clearing House, 47*, 319-20.
- Rudman, H. C. (1977). The standardized test flap. *Phi Delta Kappan, 59*, 179-85.
- Rudman, H. C. (1987). Classroom instruction and tests: What do we really know about the link? *NASSP Bulletin, 71*, 3-22.
- Rupley, W. H. (1973). Standardized tests: Selection and interpretation: ERIC/RCS report. *Reading Teacher, 26*, 752-60.
- Salmon-Cox, L. (1981). Teachers and standardized tests: What's really happening? *Phi Delta Kappan, 62*, 631-4.
- SAS Institute Inc. (1990). *SAS/STAT user's guide, version 6* (4th ed.). Cary, NC: SAS Institute, Inc.
- Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*, 3-6.
- Schmid, C. F., & Schmid, S. E. (1979). *Handbook of graphic presentation* (2nd ed). New York: John Wiley & Sons.
- Schulte A. C., & Borich, G. D. (1988). False confidence in intervals: Inaccuracies in reporting confidence intervals. *Psychology in the Schools, 25*, 405-12.
- Singer, H., & Dreher, M. J. (1983). Attitudes towards testing and test results of reading achievement. *Journal of Reading Behavior, 15*, 19-32.
- Sprinthall, N. A. (1967). Test interpretation: Some problems and a proposal. *Vocational Guidance Quarterly, 15*, 248-56.

- Sproull, L., & Zubrow, D. (1981). Standardized testing from the administrative perspective. *Phi Delta Kappan*, 62, 628-31.
- Stetz, F. P., & Beck, M. (1981). Attitudes toward standardized tests: Students, teachers, and measurement specialists. *Measurement in Education*, 12, 1-11.
- Stiggins, R. J. (1991a). Assessment literacy. *Phi Delta Kappan*, 72, 534-9.
- Stiggins, R. J. (1991b). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, 10, 7-12.
- Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22, 271-86.
- Swain, C. L. (1982). Using test data effectively. In S. B. Anderson & L. V. Coburn (Eds.), *Academic testing and the consumer*. San Francisco, CA: Jossey-Bass.
- Tittle, C. K. (1989). Validity: Whose construction is it in the teaching and learning context? *Educational Measurement: Issues and Practice*, 8, 5-13.
- Traxler, A. E. (1960). Use of results of large-scale testing programs in instruction and guidance. *Journal of Educational Research*, 54, 59-62.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology*, 32, 191-241.
- Walker, R. N. (1968). How to understand and use test results. *Grade Teacher*, 85, 105-6.
- Wang, P. C. C. (Ed.). (1978). Graphical representation of multivariate data: Proceedings of the symposium on graphical representation of multivariate data, Naval Postgraduate School, Monterey, CA, Feb. 24, 1978. New York: Academic Press.
- Whitehead, B., & Santee, P. (1987). Using standardized test results as an instructional guide. *Clearing House*, 61, 57-9.
- Williams, R. L. (1971). Abuses and misuses in testing black children. *Counseling Psychologist*, 2, 62-77.

