DOCUMENT RESUME

ED 274 710 TM 860 589

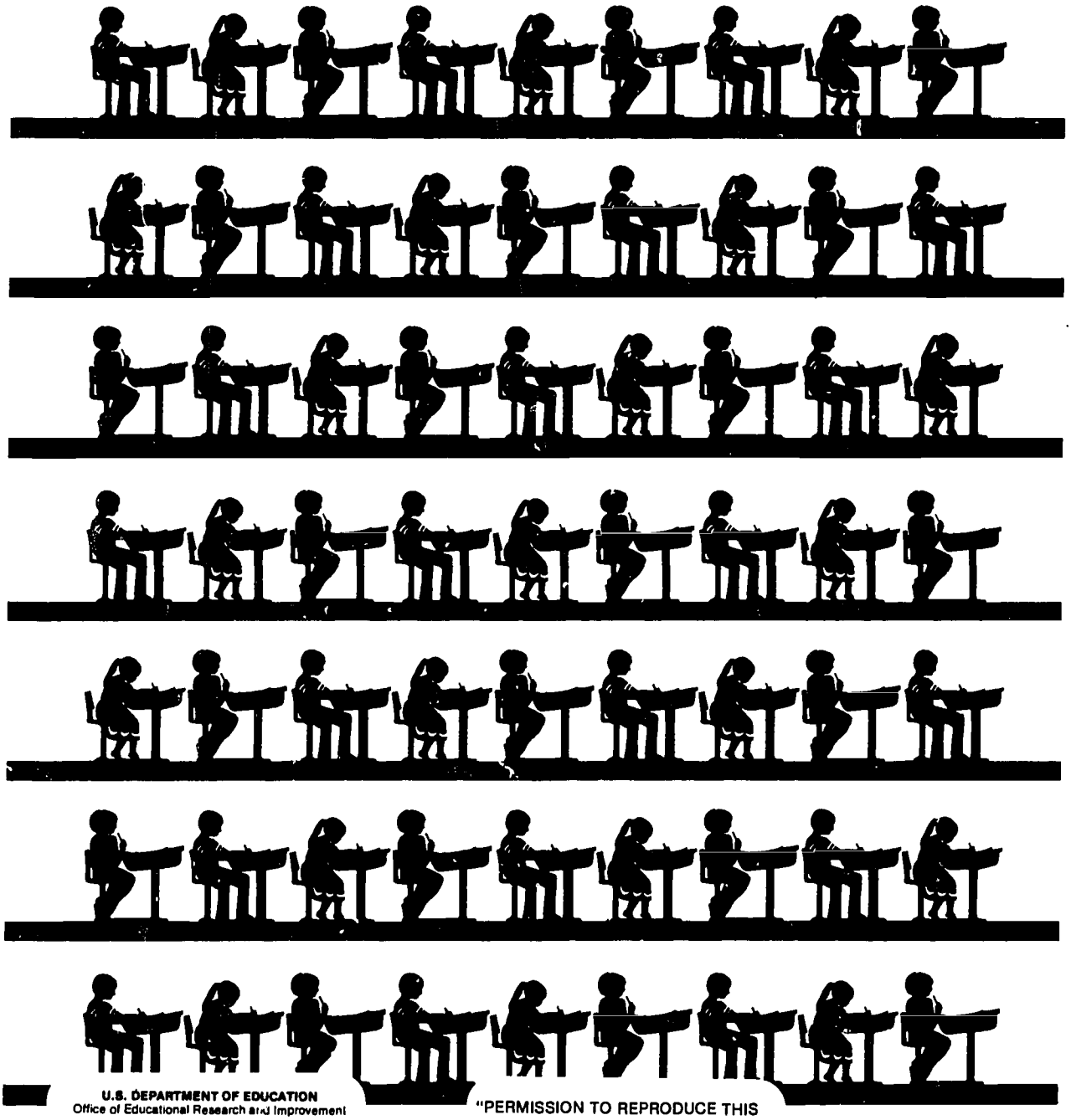
TITLE	Assessment Handbook: A Practical Guide for Assessing Alaska's Students.
INSTITUTION	Alaska State Dept. of Education, Juneau.
PUB DATE	86
NOTE	74p.
PUB TYPE	Tests/Evaluation Instruments (160) -- Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Educational Assessment; Educational Objectives; *Educational Testing; Program Costs; Recordkeeping; School Districts; Standardized Tests; State Programs; *Testing Programs; *Test Interpretation; *Test Selection
IDENTIFIERS	Alaska; *Curriculum Alignment

ABSTRACT

This handbook is designed to serve as a resource for Alaskan district-level administrators in charge of planning, implementing and monitoring districtwide student assessment programs. The information is divided into 10 chapters, including: (1) aligning assessment with curriculum and instruction; (2) planning an assessment program; (3) administering assessment programs; (4) involving constituent groups in assessment decision making; (5) selecting appropriate tests; (6) interpreting and using test results; (7) testing costs; (8) reporting assessment results; (9) keeping records; and (10) integrating statewide assessment with local assessment programs. Each chapter acts as a separate document with several articles relating to the overall topic. Nine chapters have separate sheets containing checklists and questionnaires. Additional references are listed at the end of each chapter. (JAZ)

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Alaska Department of Education

The *Assessment Handbook* was produced for Alaska's educators by the:
Alaska Department of Education
Harold Raynolds, Jr., Commissioner of Education

Bob Silverman and Al Hazelton were the Department's project supervisors. Dick Luther is the Director of the Division of Educational Program Support, which sponsored the project.

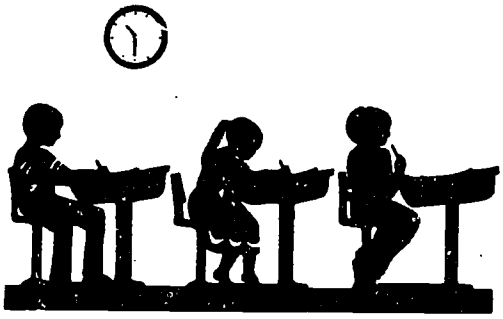
This handbook was developed with the assistance of Interwest Applied Research of Portland, Oregon. Interwest's work was supervised by Evelyn Brzezinski and Michael Hiscox, with the assistance of Elaine Lindheim, Vicki Spandel, and Frank Womer.

Graphic design by JoEllyn Loehr, Loehr Design
Typesetting by Interwest Graphics
Printing by J. Y. Hollingsworth Co.

Construction of Assessment Handbook
February 1986

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Introduction

Why Develop an Assessment Handbook?

Why do we need this handbook? What do I need to know about assessment that I don't already understand? For that matter, why should I be concerned with the issue of assessment at all? In this section we'll give an overview of the handbook which will help you to answer these questions.

Good assessment is essential to effective instructional planning, diagnosis of students' strengths and weaknesses, proper placement of students in specialized programs, and evaluation of curriculum, student achievement and educational programs. However, many of those responsible for selecting, using, and interpreting tests and their results have had little formal training or guidance in assessment.

To complicate the issue, today's administrators face a difficult educational paradox. With the growing demand for educational accountability has come increasing pressure to test. Let's find out, the public has said, whether schools are really doing their job. Yet at the same time, tests and test administration procedures have fallen under ever more critical scrutiny. Are they really measuring anything important? Are they too difficult? Too easy? Biased? Are teachers teaching to the test? Should they be? Are results being used correctly and effectively? And are the costs of testing justified?

The Need for a Practical Resource

Given the pressures for accountability and the recognized need for accurate information on students performance, how can educational administrators set realistic goals for testing and ensure that they have the proper resources to do the job? How can administrators meet the informational needs of all the groups they serve in a defensible, cost effective manner?

Administrators who have sought help in answering these questions have frequently been frustrated by the limited resources available. Professional assistance can be expensive and difficult to arrange. Textbooks are generally technical, weighted down with statistical tables and jargon virtually guaranteed to intimidate and confuse anyone who is not already a testing specialist.

The Alaska Department of Education, recognizing the need for practical, useful information, has sponsored development of this handbook. It is a nontechnical guide, designed to meet the needs of the educational administrator who wants understandable, concise, easily accessible information relating to test selection, development, administration, and interpretation. It is assumed that the primary audience for the handbook is district level administrators with responsibility for planning, conducting, and reporting the results of assessment. However, many of the concepts presented may also be of interest to school level administrators and teachers who play a role in local assessments.

The purpose of the handbook is not to train users as testing specialists; it is rather to make them comfortable with the concepts they need to understand to conduct valid, reliable assessments. You do not need a background in testing to use this handbook. You need only a serious interest in exploring the advantages and limitations of various kinds of tests, and in using tests more effectively to improve education.

Defining Assessment

The focus of this handbook is assessment. A simple statement--but what, in fact, do we mean when we use the term assessment? A simple classroom test? An elaborate standardized test series? Assessment, testing and evaluation: Are they pretty much synonymous? Or are there important differences?

The terms assessment, test and evaluation are frequently used interchangeably, but in fact there are important differences. Test is the narrowest of the terms. It denotes the presentation of a specific set of questions to be answered, a task to be performed, or a problem to be resolved. A test is tangible (or at least observable) and structured, and it can be administered within a relatively limited period of time.

Assessment is much more encompassing. While testing is a part of assessment, it is but one approach toward measuring significant characteristics about individuals or groups. Other valuable information might be gathered through informal rating scales, observations of various types, individual interviews, or reviews of a student's background or previous performance. All these methods for gathering data should be considered important components of assessment. In addition, the term assessment is often used to refer to a planned program of testing. This latter meaning of assessment is the one embraced in this handbook and so on the following pages "assessment" and "testing" are both used to imply planned programs of testing rather than single administrations of individual tests. A district's assessment program might include plans for administering anywhere from one to a dozen or more tests over some extended period. Used in this sense, the term includes, by implication, test planning, design and administration.

Evaluation, as the word itself suggests, refers to making a value judgment about the implications of assessment data. While assessment involves obtaining performance data through a variety of means, evaluation goes a step further--interpreting that data from an informed perspective. Although the handbook includes some material on interpretation of assessment results, other facets of evaluation are not included here.

In summary, the purpose of testing is to provide one isolated glimpse--analogous to taking a picture with a camera--of how a student or group of students is performing at a specific time with respect to specific skills. The purpose of assessment is to provide more comprehensive data on student performance through several administrations of the same test or equivalent tests, through administration of test batteries, or through various data gathering approaches. And the purpose of evaluation is to make value judgments about the results provided through assessment.

A word of caution here: Testing, assessment and evaluation are strongly interdependent; the quality of one affects the quality of the others. Good tests comprising sound items based on curriculum-related objectives strengthen assessment; and well planned assessment, in turn, increases the probability of accurate evaluation by providing sufficient and valid data.

What the Handbook Contains...

As noted, the focus of this handbook is assessment. Because tests are such an integral part of assessment programs, they will be discussed in some detail. Other forms of assessment will be mentioned as well, though with less emphasis. But evaluation will not be covered in the handbook. Good assessment procedures are a necessary--but not sufficient--requirement for good evaluation. Those other necessary requirements are beyond the scope of the handbook.

The information which is included in the handbook is divided into ten chapters, as follows:

1. aligning assessment with curriculum and instruction
2. planning assessment programs
3. administering assessment programs
4. involving constituent groups in assessment decision making
5. selecting appropriate tests
6. interpreting and using results
7. testing costs
8. reporting assessment results
9. keeping records
10. integrating statewide assessment with local assessment programs

Each chapter is a standalone document with several articles related to the overall topic. References for further information on the topic are included on page 4 of each chapter. Pages A and B in each chapter contain checklists, graphics and other aids that district administrators might find appropriate for reproduc-

tion and use with other district staff.

The titles of all articles are listed in the shaded box on the first page of each chapter. What if your concern doesn't seem to be covered in the titles? It may still be included in the handbook; perhaps it's listed under a title different from what you expected. Check the Glossary/Index at the end of the handbook. There you'll find definitions of all major terms and concepts listed with chapter references. The index allows you to use the handbook much the same way you'd use an encyclopedia to look up any topic of interest.

And What's Been Left Out...

Now that we've shown you the kinds of issues addressed in this handbook, it's time to mention briefly what is not included. "Testing" is a large umbrella (and "assessment" is even bigger) covering everything from the classroom spelling test to College Board exams. Our purpose in this handbook is to address the major issues related to the assessment of student achievement: the measurement of a student's knowledge or proficiency in some specific content area. Measures of intelligence or aptitude are not covered here, for several reasons.

First, such tests merit special consideration, and in-depth coverage of the production and use of such tests would require a book in itself. Second, aptitude testing (or IQ testing) is a highly controversial issue. Current thinking increasingly suggests that aptitude is not a fixed or constant characteristic, but subject to change in response to many factors, such as instruction and home environment. The potential for abuse and inappropriate use of such data is great, and rigorous special training is generally required to interpret these tests correctly. And third, but equally important, the primary priority for most districts is proficiency testing rather than aptitude testing.

Similarly, individually administered assessment measures are not discussed in the handbook. Rather, all comments relate to group administered instruments. This decision has been reached for much the same reasons that aptitude measures are not discussed--namely, that (1) specialists (e.g., counselors, psychologists) are required to interpret these test results and (2) by far the major part of any district's assessment program is based on group administered achievement tests.

Finally, the handbook does not attempt to cover the specifics of testing to meet federal and state requirements. There is, to be sure, a lot of testing done to meet the regulations of various programs. In fact, the funding itself is often contingent on providing appropriate test results to the funding agency. But the theory and requirements of that type of testing are too specific to be appropriate in this general handbook. In addition, most of the regulatory agencies have already provided some sort of testing support specific to their requirements. For example, for over ten years, the federal government has spent several million dollars a year to assist people with testing for Title I (now Chapter 1), special education, migrant education and the like. Such assistance is far greater than can be provided through this handbook. So while you may find occasional references to the role testing plays in these programs, you'll have to use other resources to get the details of these mandatory testing programs.

The handbook authors have made every effort to address the issues on which administrators most need information in order to do their jobs well. We hope you'll find both the selection of topics and the discussions appropriate and useful. After you've had an opportunity to use the handbook for a short while, we'll be sending you a feedback form to ask what you think of the handbook and what topics should be included in handbook updates. In fact, if you learn that important topics are missing, we hope you'll take the time to let us know. Write the Division of Educational Program Support at the Department of Education, Pouch F, Juneau, Alaska 99811 or call (907)465-2900.

The People Who Made the Project Happen

One of the earliest tasks of the handbook development was the formation of a ten-member Assessment Handbook Advisory Group to assist the Department and its contractor in making decisions about content, format and appropriate audiences. The group comprises five district staff (all of whom serve as District Testing Representatives for the Alaska Statewide Assessment), a member of the State Board of Education, two representatives of Alaska's professional educators associations and two staff of the Department of Education. The Advisory Group members are:

Wally Berard, North Slope

Dave Dossett, Southeast Island

Mary Francis, Fairbanks North Star

Chris Robinson, Southwest Region

Fred Stofflet, Anchorage

Sue Hull, Fairbanks (representing the State Board of Education)

Jeff Ivey, Anchorage (representing the Alaska Council of School Administrators)

Gayle Pierce, Anchorage (representing NEA-Alaska)

Myra Howe, Department of Education (special education)

Ed Obie, Department of Education (ECIA Chapter 1)

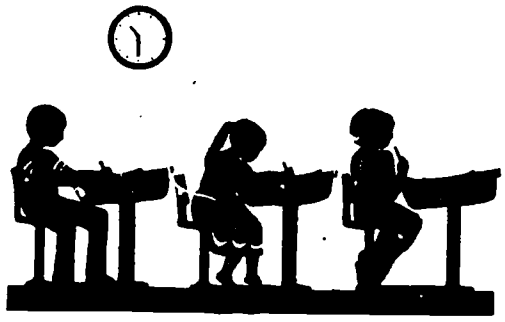
It is no overstatement to say that this handbook could not have been developed without the dedicated efforts of the members of the Advisory Group. The Department extends its deepest appreciation.

Bob Silverman and Al Hazelton of the Department's Division of Educational Program Support were the project's monitors, under the supervision of Dick Luther, Director of the division. Evelyn Brzezinski was project director for Interwest Applied Research, the contractor which assisted the Department with the handbook's development and production.

1

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Chapter 1

Alignment: A Definition

Alignment, simply put, means matching. It has numerous applications within education, so that depending on context, educators may use the word alignment to refer to a match between

- curriculum at one level and curriculum at another level,
- course content and testing,
- enabling or process skills and outcome skills, course goals and course content,
- testing and remediation (or enrichment) programs, or
- program administration and outcomes.

In the broadest sense, alignment means coordination among all the elements of instruction, curriculum and testing. When alignment is functioning well, the cyclical transitions from planning to instruction, to testing, remediation and enrichment, to evaluation, and then back to planning are all smooth. Everything works together.

Many factors affect the extent to which alignment is operating within a district. But perhaps more than any others, the following two indicators suggest the extent to which alignment has been successfully achieved:

1. Communication among educators and administrators at all levels within the district is open and functional. That is, communications channels are purposefully used to support or increase alignment.
2. Educational goals and objectives exist in writing, are coordinated across grade levels, and are well known to and supported by educators and administrators at all levels.

Of course, just because goals and objectives exist in writing does not necessarily mean that teachers are teaching those things, nor that tests are measuring what teachers are teaching. Alignment must be monitored at the classroom and building level to ensure that it is occurring. But point 2 above must have occurred at some time for building-level alignment to make sense. (By the way, a pilot effort will be underwa during the 1985-86 school year to test what alignment means at the building level. The pilot effort is being conducted under the auspices of the Alaska State Leadership Academy.)

Our primary interest in this series, of course, is the integration of assessment procedures with curriculum and instruction. But because the three areas are so intertwined, this chapter will discuss the overall issue of alignment.

Steps Toward Improving Alignment

Once you've identified alignment as a goal worth striving for, how do you go about achieving it? First, it's important to recognize that alignment is a dynamic process. Therefore, it's very hard for a district to say it has "achieved alignment." It really makes more sense to measure alignment along a continuum, realizing that it could probably always be improved, but that it certainly operates better under certain circumstances than under others.

Given that understanding, here (briefly) are the steps toward improving alignment:

1. **Review and revise (as necessary) major educational goals** across the curriculum. Be sure there is consistency among grade levels, an ordered progression from one level to the next, and appropriate tasks given the grade level.
2. **Review objectives** to make sure they match what the goals claim will be achieved.
3. **Break down the objectives** to determine precisely what skills should be introduced, emphasized or reviewed at each grade level. Also determine which skills students must master at each grade level.
4. **Ensure that all tasks identified through Step 3 can be covered by instruction.** Further, ensure that performance on these tasks can be measured through testing or other viable means (e.g., classroom observation). Consider asking those working on Step 3 to write sample test items (where appropriate) to be sure that objectives are neither too broad (in which case a variety of very different types of test items could be written) nor too narrow (in which case only one or two items could be written).
5. **Devise a plan (including assignment of responsibility) for the selection of textbooks and other materials** that will support skill development in specified areas.
6. **Outline a long-range plan for testing** that will satisfy the district's specific needs.
7. **Establish a procedure for test selection** and/or development. Be sure the procedure emphasizes criteria that foster alignment (i.e., a match between test content and curriculum and instructional content). Assign responsibilities for test selection, development and review.
8. **Review current instructional plans** and test scores to determine which curriculum goals are being addressed well, and which are being addressed poorly or not at all.
9. **Recommend new instructional strategies** that will support alignment (e.g., integrating science and math instruction so that students might develop math skills by working on science problems such as fisheries management).
10. **Survey teachers to determine inservice needs.** Do they need more information and skills related to testing? Instructional management? Team teaching or teaching across the curriculum? Designing measurable objectives?.
11. **Design inservice** based on the findings of Step 10.
12. **Provide a forum** (or more than one, if desirable) through which representatives of various groups (elementary and secondary teachers, administrators, parents and so on) can share their perceptions about instruction and testing at each grade level.
13. **Review proficiency standards** (if they apply to your district).
14. **Ensure that the testing program is sufficient to measure compliance** with proficiency standards (if they apply in your district); ensure that the instructional program is strong enough to give students the skills and knowledge they need to meet the standards.
15. **Review remedial and enrichment programs** in the district. Are placement procedures consistent across the district? Are those placement procedures directly related to district instructional and testing programs?

PRACTICAL TIP

You're not the first person to wrestle with the issue of alignment. In fact, chances are good that a neighboring district has developed an alignment procedure that might be a good starting point. If so, they will have already tried the procedures and will know what works and what doesn't. You can adopt the good and correct the bad--with much less effort than would have been necessary had you started from scratch.

16. **Make discussion of alignment an inherent part of future planning** in order to keep alignment a high priority. Ensure that all perspectives (those of teachers, principals, district administrators, content specialists, special education staff, parents, students, community representatives and others) can be heard and considered.

No one expects that a district could go through these steps for all curriculum areas in a year. But just as a district should develop a multi-year plan for devising a good testing program, so should it develop a plan for achieving ever-improving degrees of curriculum, instruction, and testing alignment. It may take several years, but the advantages for students and educators alike are well worth the effort.

Special Issues To Be Resolved

Regardless of how successfully or smoothly a district may be handling its alignment efforts, certain problems can arise. Several common problems are described below. How many apply to your district?

- **Maintaining the autonomy of each level in the educational system.** What if the elementary teachers see themselves as very different from the secondary teachers in philosophy or approach? Do secondary staff have the right to dictate to elementary staff what should be taught or emphasized? While consistency of curriculum across levels is critical, the idea of one level dictating to or directing another runs counter to the whole spirit of cooperation on which efficient alignment depends.

  What to do: Ensure that no single level takes the lead in setting educational goals or priorities. Provide a forum for discussion among representatives at all levels. Ensure that educational goals and curriculum reflect the areas of emphasis that educators from all levels view as critical.

- **Achieving alignment between life skills and academic skills.** Elementary educators may feel they have little or no role to play in development of life skills. On the other hand, secondary teachers may counter that life skills are essential to students' effective functioning in everyday life, and that achievement of such skills demands everyone's support.

  What to do: Determine first the extent to which the teaching of life skills has support in the district--from administrators, teachers, content specialists and parents. If it is truly valued, analyze the skills involved to identify critical enabling skills. Then determine at what grade levels these should be introduced, stressed and mastered by students. Bring elementary and secondary educators together not only to complete the skill level analysis, but also to exchange their views on the whole issue of life skills.

- **Ensuring that what is "covered" in a course matches what is "taught".** Suppose test questions seem to reflect course content, yet neither matches very well with what the teacher presents in class?

  What to do: First, make sure that teachers who want it have access to inservice in designing tests and relating tests to instruction. Second, set a district policy on testing for specific purposes to eliminate or minimize unnecessary testing. Third, time classroom observations so that the match between instruction and testing procedures can be checked occasionally. Finally, ensure that teachers have both time and ways to communicate with one another about which instructional procedures are particularly effective in getting across the district's curriculum objectives. (Note: A good match between instruction and test content can be extremely difficult to determine for all but the most obvious recall questions. The advice of a curriculum specialist or test developer can be very valuable in reviewing tests and designing inservice.)

- **Ensuring that testing plays a realistic role.** Not all important educational outcomes are measured through tests. Therefore it follows that good alignment does NOT demand the formal testing of everything in the curriculum. Misunderstanding of this concept can lead to overtesting which, in turn, leads to other problems such as scheduling conflicts, increased student anxiety, and staff resistance.

  What to do: Make sure that policies regarding alignment do not overemphasize testing; the match between curriculum and instruction is just as important as the match between curriculum and testing. Be clear in stating how objectives should be selected for testing. Use inservice to emphasize other valid ways of measuring students' competence, including careful classroom observation.

PRACTICAL TIP

Does the top administration understand and support alignment efforts? A lot of work can be wasted if your alignment efforts don't have consistent support.

A Summary of Advantages

The advantages to alignment may already be evident to you. Nevertheless, a brief summary seems in order. The primary advantages are these:

- Improved communication,
- Better problem solving through coordinated effort,
- Time savings,
- More efficient use of content specialists or consultants,
- Improved instruction (through consistency),
- Improved services to bilingual students, transfer students, and other special student groups,
- More efficient use of resources (because one program may serve two purposes, or meet the needs of two or more groups),
- Improved testing practices,
- Better justification of educational practices and expenditures, and
- Increased satisfaction and productivity as a result of a focused, consistent effort.

continued, over

Advantages of Alignment, continued . . .

While these advantages sound good in the abstract, what do they mean in terms of having a positive impact on students? When curriculum, instruction and assessment are aligned, there is a greater chance that a school's mission will be accomplished. Since that mission is, presumably, based on student needs, it follows that it is students who will be the prime beneficiaries of the alignment process.

Strange as it may seem, the idea of alignment among curriculum, instruction and testing is relatively new, and is evolving new dimensions as we recognize its importance to quality education. Testing what has not been taught, for example, is often thought of today as a violation of ethics and may even be subject to legal proceedings. But only a few years ago, it would more likely have been viewed as a violation of logic--and prior to that, might not have seemed a topic worthy of discussion at all.

As educators, however, we have come to see the numerous advantages in building tests that are consistent with the educational objectives of the school and its community. As those goals inevitably change, our testing practices must change to keep pace. Indeed, we are obligated to adopt a constant questioning outlook, following the advice of a noted measurement specialist, the late Robert Ebel: "It is occasionally useful to ask of any subject of study or method of instruction the simple question *'Why?'* and to insist on an answer that makes sense."

References

Fisher, Thomas H. (1983, Winter). Implementing an Instructional Validity Study of the Florida High School Graduation Test, Educational Measurement: Issues and Practice, 2, 8-9.

Describes the plan used by the Florida Department of Education to comply with a court ruling that their high school graduation test must be directly related to the curriculum and to instruction in Florida schools. The plan relies primarily on teacher and principal judgments of the match of test to curriculum.

Gronlund, Norman E. (1981). Improving Learning and Instruction. Chapter 18 in Measurement and Evaluation in Teaching, Macmillan, 483-508.

A discussion of how testing and evaluation should fit into classroom instruction. Includes how tests can help to clarify instructional objectives, assess learners' needs, monitor objectives, diagnose problems and evaluate course outcomes. Also discusses mastery learning and comments on its role in providing instructional staff with diagnostic information.

Jolly, S. Jean and Gramenz, Gary W. (1984, Fall). Customizing a Norm-Referenced Achievement Test to Achieve Curricular Validity: A Case Study. Educational Measurement: Issues and Practice, 3, 16-18.

Describes how the Palm Beach County, Florida school system matched test questions from the national standardized test they were using to their own local objectives of instruction. The national test was rescored using only items matching Palm Beach objectives, and additional test questions were written to cover local objectives not covered by the national test.

Mehrens, William A. (1984, Fall). National Tests and Local Curriculum: Match or Mismatch? Educational Measurement: Issues and Practice, 3, 9-15.

A thorough review of the concerns relating to whether and/or how standardized achievement test results relate to objectives of instruction in a local school district. Discusses appropriate and inappropriate inferences that can be made from test results. For the reader who already has some background in this topic area.

Perkins, Mary R. (1982, Winter). Minimum Competency Testing: What? Why? Why Not? Educational Measurement: Issues and Practice, 1, 5-9, and 26.

Discusses the various definitions of minimum competency testing, along with 26 claimed "benefits" of MCT's and 24 potential "costs" or problems with such testing. For the person who wants an overview of pros and cons of minimum competency testing.

Checklist for Determining Alignment

Various criteria have been developed for gauging a district's alignment of curriculum, instruction and assessment. The Department of Education's Office of Curriculum Services has prepared a set of guide questions under categories such as mission statement, school board, budget, conditions for learning, and others.

The following checklist is another approach to determining alignment. Use the checklist to judge your district's progress toward achieving curriculum alignment. Circle a number beside each statement to show how true it is in your district, using the following scale:

4 = Very true in the district

3 = Somewhat true in the district

2 = Mostly untrue in the district

1 = Completely untrue in the district

0 = Don't know whether it is true in the district

Goals and Objectives

1. Clear educational goals and objectives have been established and put in writing. 0 1 2 3 4
2. The goals and objectives are coordinated across grade levels, with an ordered progression from one level to the next. 0 1 2 3 4
3. The objectives match what the goals state will be achieved. 0 1 2 3 4
4. The district's goals and objectives are known and supported by educators and administrators at all levels. 0 1 2 3 4

Instruction

5. Teachers design their instruction to match the district's objectives. 0 1 2 3 4
6. Textbooks and other materials are selected because they support the skills needed to meet the objectives. 0 1 2 3 4
7. Teachers are provided with inservice activities that support the alignment process. 0 1 2 3 4
8. The needs of special groups of students (bilingual, special education, gifted and talented, etc.) are addressed. 0 1 2 3 4

Testing

9. Achievement tests consistent with the district's goals and objectives are administered on a regular basis. 0 1 2 3 4
10. Tests are matched to course content and materials. 0 1 2 3 4

continued, over

11. Tests are matched to classroom instruction.	0	1	2	3	4
12. Test results are used to evaluate which goals are being achieved and which are not.	0	1	2	3	4

Communication

13. Open and functional communication exists among educators at all levels within the district.	0	1	2	3	4
14. Forums are provided for interested groups (parents, teachers, community members, for example) to share their ideas about curriculum, instruction and testing.	0	1	2	3	4
15. Teachers at one level or in one content area work closely with their colleagues at other levels and in other subject areas to achieve common educational goals.	0	1	2	3	4

When you have completed assigning 0-4 ratings to each of the criteria, go back and find all the criteria that have ratings of **0, 1 or 2**. Circle the number of these "troublesome" criteria below.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15

Now it's time for an honest appraisal. Can you improve the situation associated with the problem criteria you circled above? Talk with others as you decide which of the problems can be eliminated and which will be present thoughout the alignment effort. Circle the appropriate number below for all those criteria that **cannot be met**, even with special effort.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15

You may not have had to circle any number, but more likely, there will still be one or two important components of the alignment effort that just won't fall into place. You can learn to live with these shortcomings, but it is important for everyone associated with the alignment effort to know that these problems exist.

2

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students

Chapter 2

The First Step: Determining Purposes for Assessment

The underlying purpose in conducting any assessment is to obtain the information needed for thoughtful, intelligent decision making. Therefore, it is critical to know at the outset what decisions must be made so that one can design an appropriate assessment. Good assessment planning begins with defining these decisions--that is, with defining the purposes for the assessment. To test as a matter of tradition or convenience is totally inappropriate.

Generally speaking, assessment results feed into one of two kinds of educational decisions--instructional management or programmatic decisions. Let's consider each individually.

Using Tests for Instructional Management

Diagnosis, placement, guidance, screening admission, and certification are all decisions related to instructional management for individual students.

Diagnosis: Results from an assessment are used to plan a specialized program that meets the needs of that student, given his or her particular strengths and weaknesses.

Placement: Test results can be used to help make decisions about which courses of study a student should enter, and at which level--advanced reading, remedial reading, or somewhere in between, for example.

Guidance: While placement involves assigning a student to a particular course, guidance matches a student with a whole program of study (such as college preparatory or vocational).

Screening/Admission: Where space is limited, not every student who applies for a course or program can be admitted. Or when educational activities are developed to be of special benefit to certain groups of students, there must be some fair and objective way of identifying the students who can best take advantage of the special programs. In both these cases, tests can be used to help determine which students should be admitted.

Certification: Tests may be used to certify minimally acceptable competence in a given area--e.g., in pass/fail courses or for promotion to the next

grade or graduation. (This latter use of tests is fraught with legal ramifications that are beyond the scope of this series. See the relevant references listed on page 4 for additional information on this topic.)

Using Tests for Programmatic Decisions

Three primary types of decisions are involved here: accountability, research/planning and evaluation. All three are concerned with assessment results from groups of students rather than from individuals.

Accountability: Citizens increasingly are questioning the nation's educators about the quality of education provided to students. One thing they have asked to see as a measure of schools' accountability is test scores. While one hopes that laypersons recognize that tests do not measure all significant aspects of schooling, there is no chance that the demand for test scores will go away.

Research/Planning: Student achievement data can pinpoint aspects of the educational system needing further investigation. Test scores can't explain why results are the way they are, but they can raise leading questions. And with answers to these questions (through research), planning for improved educational opportunities can follow.

Evaluation: Evaluation implies that judgments of worth about something are being made. Evaluation usually starts with assessment results, but it should go beyond such data before decisions based on those value judgments are made.

These eight purposes for testing, then, guide all further decisions about assessment programs. No district tests for just one of these purposes, nor can any single test possibly serve every testing purpose. But there are always priorities, and a well designed testing program can take advantage of potentially overlapping information needs. Being sure of the priorities in your district is critical if you are to develop the best assessment program for your own needs.

Test Purpose Suggests Best Test Type

Different purposes for testing demand different types of tests. The most common classification of a test is as either norm referenced or criterion referenced. What are the differences between these two?

In a norm referenced test (NRT), a student's score is interpreted by comparing it to the performance of other students. Stated another way, it is relatively unimportant exactly what content a student actually knows as shown by an NRT; whether the student knows more or less than other students taking the same test is the important thing.

In a criterion referenced test (CRT), on the other hand, a student's performance is judged according to some specified standard. How much a student knows is the important thing, not whether he or she knows more or less than the other students tested.

A logical question would be "Isn't there some overlap here? Couldn't a test be criterion referenced and norm referenced at the same time?" The answer is yes. The items on an NRT can be matched to district objectives so that criterion referenced interpretations can be made--see Chapter 5 of the *Assessment Handbook* for more information on this topic. And conversely, CRT results can be ranked to see how students compare with each other.

PRACTICAL TIP

You won't always find clear-cut references to norm- and criterion-referenced tests in publishers' sales materials. After all, you'll be happier with your test if you think it is "all things to all people." But the way an NRT is constructed is fundamentally different than a CRT. It's handy to get some criterion-referenced information from an NRT, but don't set your expectations too high.

The following chart shows the testing purposes described in the previous article and the most appropriate test type for each purpose. In some cases, both NRTs and CRTs are appropriate for a given purpose, depending on the specific question asked. For example, when accountability is the purpose for testing, an NRT is appropriate if the question being asked is "How are our students achieving compared to their counterparts elsewhere in the country?" while a CRT is appropriate if the question is "Are our students learning what we say we're teaching them?"

In summary, remember that neither testing approach (norm referenced or criterion referenced) is inherently better than the other. It makes no sense to ask "Which approach is preferable?" but rather to ask "Which approach is preferable for a given purpose?"

Purpose	Most Appropriate Test Type
Diagnosis	Criterion referenced
Placement	Criterion referenced
Guidance	Norm referenced
Screening	Norm referenced
Certification	Criterion referenced
Accountability	Norm or criterion referenced
Research	Norm or criterion referenced
Evaluation	Norm or criterion referenced

Deciding When to Test

Once the testing population is identified (see p. 4), the next questions to answer involve when to test.

How Often Should Tests Be Given?

Often the primary factor affecting this decision is cost. It is not, however, the only factor that should be considered. Of equal importance is the underlying purpose of the assessment. For example:

- Diagnostic testing requires adequate time for analysis of results and the planning of remediation.
- Testing for formative evaluation (conducted while a program is still in developmental stages) must allow enough time for data analysis, planning and implementation of new procedures.
- Testing for summative evaluation (conducted when a program is completed and decisions about continuation are being made) must be delayed long enough for program revisions to have had major impact.

Once decisions on ideal timing are worked out, the question "Is the ideal plan affordable?" must be raised. Perhaps every-year testing is desirable but just not feasible. Some compromise--perhaps involving alternating grade levels--can usually be worked out.

PRACTICAL TIP

It's common for an Alaska district to change testing specialists. Be certain that the person in charge puts important testing information in a file (or better yet, a notebook) that can be passed on to a successor.

What Time of Year?

Beginning of year testing theoretically allows time for building in remediation or trying new approaches with students identified as needing help. But it can be a real challenge to plan and implement new programs when instructors' time is taken up with planning and delivering day-to-day instruction.

The major advantage to spring testing is that students have the benefit of extensive instruction to master skills covered by the test. Also, students who are identified as needing assistance can take advantage of summer programs. In addition, the summer break allows time to structure new programs, or make desired revisions in existing ones.

There are disadvantages to spring testing, though: scores often aren't returned before the end of school, it is sometimes hard to maintain student motivation as summer approaches, and many students leave and enter the district during the summer (meaning that some test scores are useless while other students don't have any data).

Regardless of the time of year selected for testing, if standardized norm referenced tests are used it is important to pay attention to the test's norm dates. Tables of norms are developed to be used during certain periods of the year; if a district does not test during those periods, comparison of district performance with the norms tables is not valid.

Days of Week/Time of Day

While individual circumstances should always be considered, experience suggests these guidelines be followed:

- Don't test on Mondays or Fridays; more students tend to be absent those days.
- Don't test right before or after a holiday; this is another time when more students are likely to be absent.
- Test in the morning; students are often more alert then.
- Don't test right before or after lunch, recess or special activities; students' concentration is diminished then.
- Don't cram all testing into the shortest possible period; instead spread long test administrations over several days.
- Test all students taking the same test in the same way, on the same day, at the same time of day.

What Content Should Be Tested?

Most assessment programs address basic skills, which include reading, mathematics and language arts. More and more districts are adding social studies and science to the list of content areas tested. And some districts set as a goal the assessment of all content areas when they embark upon criterion referenced test development or selection.

Unfortunately, because testing time and resources are limited, assessment of many content areas generally means less than adequate coverage of most topics within those content areas. More can be gained by thoroughly assessing only one or two content areas, thereby obtaining a more valid picture of students' overall skills in the subject area selected.

Deciding what to test depends on the purpose for testing. If the purpose is curriculum evaluation, for example, districts might consider assessing content areas on a cycle (e.g., reading and language arts in year 1, science and math in year 2, social studies in year 3, arts and vocational subjects in year 4, then back to reading and language arts in year 5). This could be arranged so that test results feed into regularly scheduled curriculum and textbook reviews.

Or, if the purpose for testing is diagnosis, the district should determine just what resources are available to address deficient skills before deciding which subjects to test. If resources (staff and materials) will allow remedial assistance to be given in just elementary reading and mathematics, then it makes no sense to acquire or produce diagnostic tests in language arts or science or higher level math.

Which Students Should Be Tested?

What factors should be considered in planning whom to test? In identifying the target population for testing, there are two primary questions to keep in mind: the proportion of students to be tested and the grade levels that should be tested. Let's consider the questions one at a time.

How Many Students?

When the decision(s) to be made based on test scores affect broad groups of students rather than individuals, sampling is desirable. Unfortunately, the small size of most Alaska districts makes statistically representative sampling impossible in all but a few of the state's districts.

But there are other ways to sample besides selecting a subgroup of students. Content can be sampled; reading can be tested in one year, math in another and so on. Grades can also be sampled. While most Alaska districts test every grade every year, remember that testing should be occurring because decisions are being made based on the results--either for individual students or for educational programs. Those decisions may not really require testing every student in every subject every year.

What Grade Levels?

How do you decide which grade levels should be tested? Here are some factors to consider.

> How does the test content correspond to the curriculum?
>
> What is the structure of the school system by grade level?
>
> What resources will be available to help students who need assistance?

The first consideration requires that decision makers know both whether the tested skills are included in the district curriculum and when students are expected to master them. The second suggests that "milestone" grades--grades where there are changes in the schools that students attend, where students go from a single instructor to many, and so on--are particularly important to decision making and, thus, should be tested.

Equally important, testing cannot be separated from the issue of what to do with students whose performance is in some way unsatisfactory. The resources necessary for administering a test are only a fraction of what is required to make instructional improvements based on the results. It is not usually a good idea to use limited resources to test when nothing will be done with the data.

References

Anderson, Beverly (1982). Test Use Today in Elementary and Secondary Schools. A chapter in Ability Testing: Uses, Consequences and Controversies, Part II, National Academy Press, 232-285.

> Presents the results of a survey of testing programs and practices in selected school districts. Provides very detailed tabular information about all aspects of school testing. Good background information for a system that wants to see what other districts do.

Citron, C. H. et al. (1983, Winter). Debra P. v. Turlington 564 F. Supp. 177. Educational Measurement: Issues and Practice. 2, 6-14.

> Six articles related to the legal issues that arose when Florida instituted a high school graduation examination program which proposed to withhold diplomas from students scoring below a cutoff.

Gronlund, Norman E. (1981). Testing Programs, Trends, and Issues, Chapter 15 in Measurement and Evaluation in Teaching, Macmillan, 404-429.

> Suggests a minimum testing program for a school system, along with its rationale. Spends more space on current trends and issues in school testing. Also covers external testing programs (SAT and ACT).

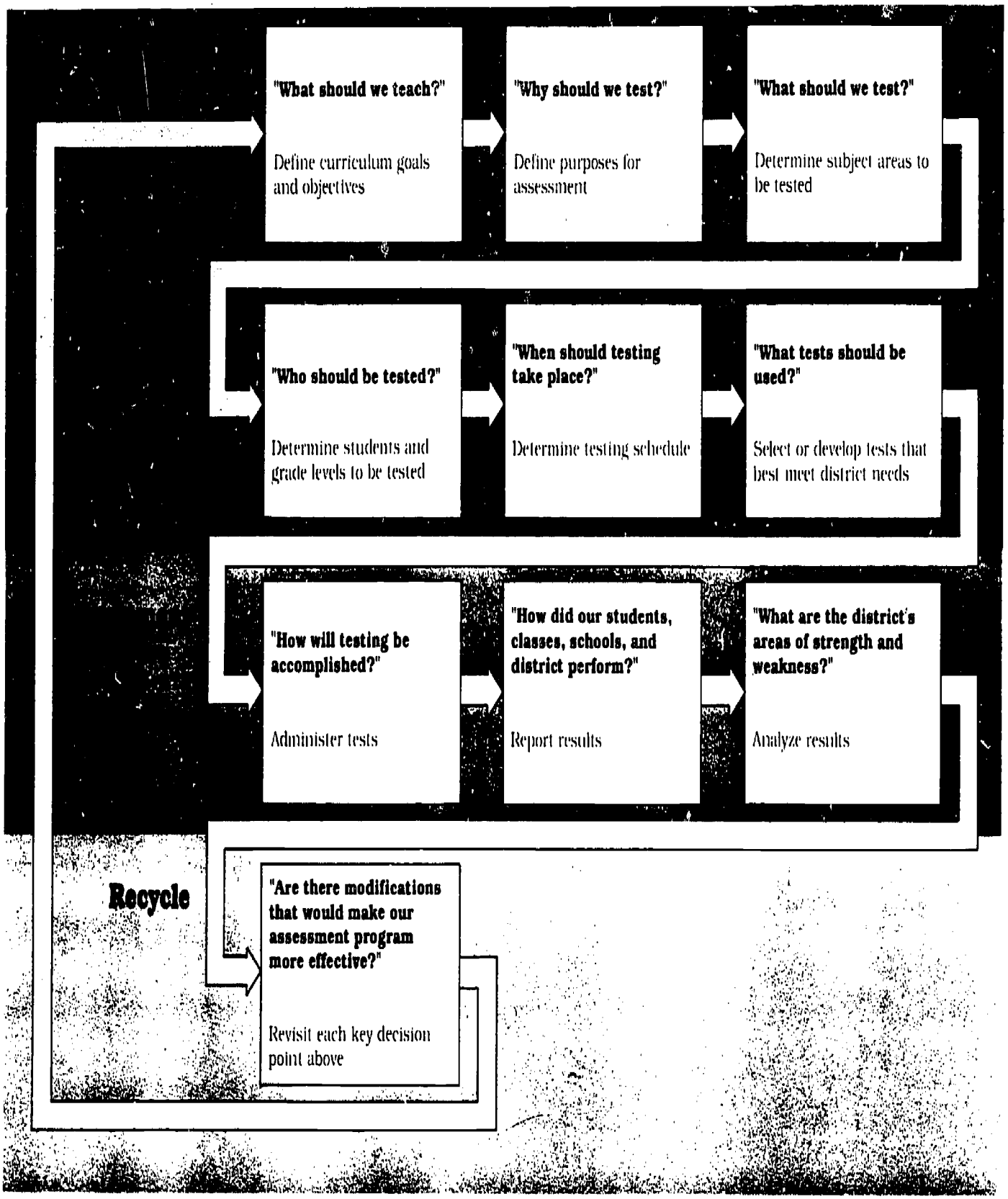
Kearney, C. Philip (1983, Fall). Uses and Abuses of Assessment and Evaluation Data by Policymakers, Educational Measurement: Issues and Practice, 2, 9-12 and 17.

> Discusses and illustrates how a testing/assessment program both in a local school district and at a statewide level can provide information useful for school administrators who have a variety of policy decisions to make. Illustrates the need for thinking about how data results are to be used before finalizing plans for a testing program.

Womer, Frank B. (1979). Uses and Abuses of Tests in Southeastern Regional Conference on Testing and Instruction, The Southeastern Regional Council for Educational Improvement, 111 Coliseum Blvd., Montgomery, Alabama 36109, 6-12.

> Concentrates on how one should and should not use test results from cognitive tests (achievement and ability). Stresses "planning" for test use along with discussing five specific things "to do" and four specific things "tc avoid" when using test results.

The Evolution of an Assessment Program



2-A

Assessment Planning Worksheet

Use this worksheet to guide the assessment planning process in your district. If your district selects several different testing purposes, you may find it easier to complete a separate worksheet for each purpose and then combine the information that results.

Test Purpose

(Check as many as apply)

- ____ **Diagnosis** (Determine students' strengths and weaknesses in specific areas)
- ____ **Certification** (Determine which students have mastered specified content)
- ____ **Placement** (Place students into appropriate level of instruction)
- ____ **Accountability** (Report the effects of an educational program)
- ____ **Guidance** (Match students with appropriate educational or vocational programs)
- ____ **Research/Planning** (Isolate educational areas needing further investigation)
- ____ **Screening/Admission** (Decide which students to select for a program)
- ____ **Evaluation** (Judge the worth of an educational program)

Test Type

(For each test purpose you have selected, indicate the type of test that will be used.)

NORM-REFERENCED TEST
(Compares examinee scores against performance of other students)

Will be needed for the following test purposes:

_____

_____

_____

CRITERION-REFERENCED TEST
(Compares examinee scores against a specified standard)

Will be needed for the following test purposes:

_____

_____

_____

Areas to be Tested

(Check as many as apply)

- ____ Reading
- ____ Language Arts
- ____ Writing
- ____ Mathematics
- ____ Social Studies
- ____ Science
- ____ Foreign Language
- ____ Fine Arts
- ____ Vocational Arts
- ____ Academic Aptitude
- ____ Attitudes (e.g., attitude toward school)
- ____ Other: _____

Grades to be Tested

(Check as many as apply)

- ____ K
- ____ 1
- ____ 2
- ____ 3
- ____ 4
- ____ 5
- ____ 6
- ____ 7
- ____ 8
- ____ 9
- ____ 10
- ____ 11
- ____ 12

Special Populations to be Tested

(Check as many as apply)

- ____ Chapter 1
- ____ Gifted
- ____ Limited-English Speaking
- ____ Special Education
- ____ Other: _____

Frequency of Testing

- ____ Twice a year
- ____ Once a year
- ____ Once every other year
- ____ Once every _____

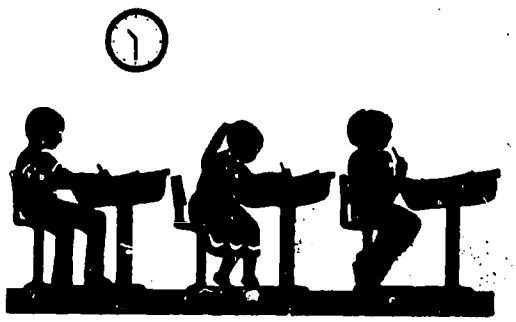
Testing Times

- ____ Fall
- ____ Winter
- ____ Spring
- ____ Summer

3

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Chapter 3

Things to Do Before, During and After Testing

Often so much district effort is devoted to designing a testing program--selecting standardized tests, developing local tests, and so on--that, once the program is designed, little additional thought is given to how the program will be administered. That oversight can have negative impact on district administrators, teachers, and students alike.

Important details must be attended to before tests are administered, when they are administered, and after they are administered. Some of these details are things that can be handled by just one person in the district; most of them, though, require a cooperative effort among administrators, teachers, students and even parents. Checklists on pages 3A and 3B show the various steps involved. The listing below explains those steps which might not be self-explanatory.

Before Testing Occurs

- Determine the number of tests, answer sheets and manuals that are needed for each grade to be tested. It is a good idea to order extra tests and answer sheets (say 5-10% more than you think are needed) so they are available in an emergency.
- Because of test security, don't stockpile standardized tests. On the other hand, you must allow plenty of time for the materials to be received in the district so they can be distributed to schools in sufficient time for the testing. Ordering materials three months prior to planned use seems a good compromise.
- As materials arrive from the printer or publisher, check to make sure that everything necessary was received. Don t just open boxes, though. Determine that sufficient quantities of all materials were received so no last-minute calls to the publisher or printer are necessary.
- As testing time approaches, organize the materials for easy distribution. Again, because of test security, it is probably not a good idea to distribute tests to schools much more than two weeks before testing. But the tests can be arranged in a locked central location so that they are ready for distribution at the appropriate time. It is probably most efficient to organize

materials (test booklets, answer sheets and administration manual) in classroom packages for each school.

When Tests Are Administered

- Be sure that the room in which testing occurs has a clock with a second hand. If such a clock isn't available, the test administrator should be provided with a stopwatch. The test administrator will then have to write the time on a blackboard at intervals so students know how long they have left to work on the test.
- Read through the administration manual and the test before the actual testing period starts. In this way, the test administrator can highlight any words in the directions that students might not understand.

PRACTICAL TIP

A class of students on testing day is neither the audience nor the time to question a testing program. Students get cues about how to react to a test from their teachers. They need to know their teacher wants the test information. If their teacher reads the directions carefully and proctors the test rigorously, students will give their best effort. But if the teacher conducts the testing in a haphazard manner, paying little attention to either the directions or the class behavior, the students will react accordingly and the test won't indicate what students know. The money spent on the testing is money down the drain.

After Testing is Finished

- Consider making copies of the answer sheets as protection against loss in the mail. It's doubtful that this should be a standard procedure for all schools. But if the school or district has a history of lost mail, it's something to consider.
- Specify any special scoring options, summary reports or data handling so it will be clear to those conducting the scoring. It is much more expensive to go back after the initial scoring and perform subsequent analyses than to do them the first time through. The decision on reports to receive should have been made at the time the tests were selected or developed.

When Scores Are Received

- Let students know how they did on the test. It's very frustrating to be asked to provide information and then never learn how the information was used. In a testing situation, that use of information translates to a score. Students should be told how they personally did on the test or, if only group results are available, how their group performed. If appropriate, use the discussion of test results as an opportunity for reteaching.

Preparing Students to Take Tests

It seems nonsensical to talk about preparing students to take tests; the best way to prepare them, after all, is to be sure that they have mastered the content areas that the test measures. But research has shown that some students are "testwise"; that is, they have certain skills which are independent of their knowledge of the subject matter being tested but which make them better at taking tests. It results in a small but consistent difference in test scores in favor of students who are testwise.

Testwiseness can be taught to students and it is not considered unethical to do so. Here is a list of some things that would pay off for students if they could learn them. The steps don't make students any "smarter," but they help ensure that students get credit for everything they do know.

- Good tests don't have trick questions; therefore, don't look for things that aren't really there.
- Choose the most correct answer, even if more than one of the choices may be partly true.
- Don't look for patterns in the answers; don't worry if there are five "A" answers in a row, for example.
- Always estimate the answer for a number problem before working it so you can see if your final result is reasonable.
- Use only the facts given in the test unless the purpose of the test is to see how many other facts you can recall. In a math story problem, for example, just use the information provided in the story, even if you know something about the situation in real life.
- Don't worry if too much or too little information is provided in the test question; sometimes questions are designed to see if you know what information to select in order to answer the question.
- Look at the questions which accompany a reading passage or story problem before reading the passage; this helps you notice the facts you need.
- Watch out for "None of the above" and "All of the above" answer choices. These choices make the items harder because you have to decide whether there is a right answer at all or whether more than one answer may be correct.
- Be sure you answer the question that's asked. A response choice might be perfectly true, but not be the answer to the question that's asked.
- Pay attention to the practice items; they show both what the test is like and how to mark your answers.
- Guess if you don't know the answer, especially if you're sure that one or more of the response choices is wrong.

- On your first pass through a test, skip the questions you find hardest. Go all the way through the test answering the easier items, then go back and work on the harder ones.
- Be very careful marking your answers on computer-read answer sheets. Be sure that marks are erased completely if you change your mind about an answer, and don't put stray marks on the sheet. Make your marks dark enough so that the computer will pick them up.
- Keep aware of the time. Make a note of what time it will be when there are ten minutes left in the testing time; save those ten minutes to review your answers.
- Get a good night's sleep before a test, and eat a good breakfast that morning. But don't drink a lot of liquids or eat a big meal right before the test is given.
- Stay calm. Tests aren't a punishment and, while it's OK to be a little anxious about them, they shouldn't make you overly tense or upset. Just try to relax, concentrate, and do your best.

What Makes a Standardized Test "Standardized"?

A standardized test is called that because it is assumed that all students who take the test are given it under the same standard administration conditions. Those conditions are the ones used when the norm group was tested. If the standard conditions aren't met when a student is tested, then the comparison of results with the norms is invalid. There are several changes in administration procedures that could invalidate results, including the following, all of which should be avoided:

1. Haphazard reading of the directions. The administration manual will tell you which directions must be read verbatim and which can be paraphrased or expanded. Pay attention to these directions. In most cases, there will be a chance to clarify instructions when the students respond to the practice items.
2. Not timing the test exactly. Allowing more time gives students an unfair advantage over the norm group. Spending less time than the norm group was allowed will make students' scores lower than they deserve to be.
3. Reading questions aloud that are meant to be read silently by the student, defining words in the test items, or explaining what the item is asking of the student. Although it is appropriate to answer procedural questions about a test, it is never appropriate to answer questions about content.
4. Translating items. If a student is in his or her first year at an English-speaking school and is used to communicating in another language, the student should probably be excused from taking the test in English because the test scores will not really be an accurate representation of what he or she knows. And translating the items into the student's native language really is not an acceptable solution to providing information about the student's skills in relation to the norm group. The only type of information that can be provided by translating test items is criterion referenced information--how well a student can perform certain skills when they are tested in his or her native language.

In summary, the information publishers give about a student's test scores is based on the assumption that directions and other testing conditions are the same as when the test was administered to the norm group. Deviating from the publisher's instructions will invalidate the information provided by the test.

Although there may be an altruistic desire to help students during the test, the test administrator should avoid the above actions. Again, the information obtained about students will be useful only if the directions are closely followed.

Teaching to the Test vs. Teaching the Test

During the second day of meetings before school starts in the fall, the district's teachers pore over copies of their standardized achievement test, which will be administered districtwide in April. Some teachers are observed copying items. In another district, teachers spend a substantial portion of a morning reviewing the objectives which are measured in their locally developed criterion referenced tests (CRTs). The first instance is clearly unethical. What about the second?

No, it's not unethical. In fact, it's something to be supported. There are two major differences in these situations that account for the fact that the second activity is commendable while the first is to be guarded against.

The first difference is that it is standardized test items that are being studied while it's CRT objectives that are being reviewed. The second difference is that the CRT objectives coincide with the district's curriculum; the tests were developed to measure that curriculum. The same statement is unlikely to be true for the standardized test.

So in reality, when the CRT objectives are being reviewed, it's actually the district's objectives that are being studied. That's what teachers are supposed to have clearly in mind. The fact that they're measured by the CRTs is incidental and clearly not unethical.

How Teachers Can Help Students

Just as there are things that students can do during a test to show their knowledge to best advantage, so are there things that a teacher can do to help students get ready to take a test. What are some of those things?

- If the test is an untimed one, make certain students have sufficient time to complete it. The Alaska Statewide Assessment Tests, for example, are untimed tests. Students should be given as much time as they need--as long as they are making progress--to complete the test.
- Alternately, if the test is a timed one, be sure to keep students apprised of how much time is left for them to complete the test. It's better to write the time on a blackboard than announce it.
- Do the practice items. If the teacher doesn't review these together with the group, many students will ignore them. But they are very important, for they show students if there is anything unusual about the way the test questions are worded and also afford students the chance to be sure they know how to mark their answers.
- Make certain that students take the test seriously. While it's unkind to play on students' anxiety by stressing the importance of the test, it's also inappropriate to underplay it. Present the testing situation with a positive attitude about both the usefulness of the activity and the students' ability to cope with it.

References

Gronlund, Norman E. (1981). Test Selection, Administration, and Use. Chapter 11 in Measurement and Evaluation in Teaching, Macmillan, 275-302.

This chapter includes a section labeled "Administration of Published Tests" which presents a concise Test Giver's Checklist. The sections on test selection and test scoring also are useful.

Hills, John R. (1981). Developing Test-Taking Skills (Chapter 7) and Finding, Choosing, and Administering Standardized Tests (Chapter 11) in Measurement and Evaluation in the Classroom. Merrill, 119-134 and 209-229.

Chapter 7 discusses test-taking skills from the point of view of the student, including testwiseness and test anxiety. In Chapter 11 there is a concise but thorough section on test administration.

Iverson, Grace (1984, Summer). Raising Test Scores, Educational Measurement: Issues and Practice, 3, 45-46.

Describes a Lansing, Michigan school district plan for helping to improve its Michigan State Assessment scores. It includes a description of working with the local newspaper to publicize the plan, implement it, and report test results.

Michigan State Board of Education (undated). A Guide to Test Taking: As Easy as 1, 2, 3. A publication of the Michigan Educational Assessment Program, Box 30008, Lansing, Michigan 48909, 36 pages.

A pamphlet for teachers who want to help their own students improve their general test-taking skills. Also includes sections specifically for students and parents. Includes sample tests and is full of practical hints.

Womer, Frank B. (1980). Preparing for an Examination--Guidelines for the Student. Chapter 1 in Review of Dental Assisting, C. V. Mosby, 1-7.

Although this chapter is aimed at students preparing to take the "Certification Examination for Dental Assistants," it covers basic principles and practices for long term improved test taking.

Administering an Assessment Program

District-Level Responsibilities

Use this checklist to keep track of district-level responsibilities for administering a successful assessment program. Space has been provided at the end of each section for you to add other activities that may be required.

BEFORE TESTING

___ Schedule testing dates for district.

___ Determine number of students to be tested.

___ Determine number of tests, answer sheets, and test manuals needed (allow 5-10% overage).

___ For commercially published tests, place order with test publisher three months prior to testing date.

___ For locally developed tests, complete printing of materials one month prior to testing date.

___ If scoring will be done locally, order or prepare scoring materials (keys, report forms, directions, etc.).

___ When test materials arrive from the publisher or printer, check them over carefully.

___ Package materials for distribution to school sites.

___ Distribute materials to schools two weeks before testing.

___ _____

___ _____

DURING TESTING

___ Be available if schools have questions or need additional materials.

___ _____

___ _____

AFTER TESTING

___ Check in materials returned from school sites.

___ Discuss testing with school-site staff to determine if there were problems or concerns.

___ If scoring will be done locally, prepare and process materials according to established procedures.

___ If scoring will be done elsewhere and your district has a history of lost mail, copy answer sheets before mailing them.

___ Bundle answer sheets according to publisher's instructions.

___ Notify publisher if any out-of-level testing has been done.

___ Specify any special scoring options, summary reports, or data handling desired.

___ _____

___ _____

WHEN SCORE REPORTS ARE RETURNED

___ Distribute test results to schools.

___ Provide teachers with training in interpreting test results to students and parents and in using test results for instructional improvement.

___ Share test results with concerned groups (parents, school board, newspaper, etc.).

___ _____

___ _____

Administering an Assessment Program

School-Site Responsibilities

Use this checklist to keep track of the school-site responsibilities for administering a successful assessment program. The checklist is divided into two parts. Part I lists activities that are the responsibility of a single school coordinator. Part II lists activities that are the responsibility of each teacher who administers tests. Spaces have been provided in each part for you to add other activities that may be required.

PART I: SCHOOL COORDINATOR

BEFORE TESTING

___ Arrange for appropriate testing rooms (adequate seating and light, comfortable temperature and ventilation, no distractions from outside).

___ Distribute test materials to teachers.

___ Provide any special test administration training needed.

___ Inform parents in advance of testing dates and offer suggestions for preparing students for the test days.

___ _____

___ _____

DURING TESTING

___ Be available if teachers have questions or need additional materials.

___ _____

___ _____

AFTER TESTING

___ Arrange for make-up testing as needed.

___ Return materials to district coordinator.

___ _____

___ _____

WHEN SCORE REPORTS ARE RETURNED

___ Distribute reports to teachers, parents and students.

___ _____

___ _____

PART II: TEST ADMINISTRATORS

BEFORE TESTING

___ Prepare students by discussing the purpose of the test and teaching "testwiseness" hints.

___ Motivate students to do their best work.

___ Have an extra supply of pencils available.

___ If testing room does not have a clock, have a stopwatch available to post time on board.

___ _____

___ _____

DURING TESTING

___ Arrange room so that all desks face the front.

___ Check that lighting, temperature and ventilation are all optimum.

___ Put "Do Not Disturb" signs on doors.

___ Make sure that each student has a test booklet, answer sheet and pencil.

___ Follow all procedures as described in administration manual.

___ Complete all practice items with class.

___ If tests are timed (e.g., many standardized tests), keep students informed of the time left to work.

___ If tests are untimed (e.g., Alaska Statewide Assessment Tests), allow students to take as much time as they need to finish their work.

___ Circulate during test to make sure that all students are following the directions and marking their answer sheets correctly.

___ Answer procedural but not content questions.

___ _____

___ _____

AFTER TESTING

___ Check answer sheets for names, completeness of other identifying information, dark marks, clean erasures, and no stray marks.

___ Return materials to school coordinator.

___ _____

___ _____

WHEN SCORE REPORTS ARE RETURNED

___ Interpret test results for students and parents.

___ Use test results for instructional improvement.

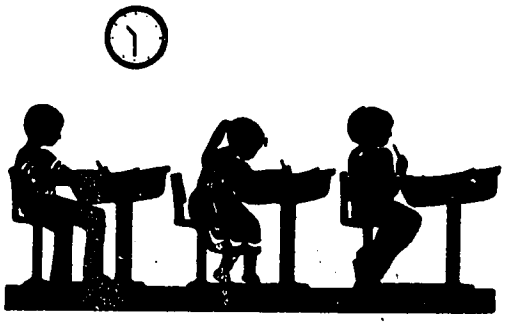
___ _____

___ _____

4

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Chapter 4

Different Groups--Different Roles

If an assessment program is to play any significant role in education, it must have the support of both the taxpaying public that funds it and the educators and administrators who work with it. The program must appear credible--that is, valid and useful--to teachers, administrators, parents and students alike. How can district administrators develop needed credibility for a program?

One major way is by involving these groups in decision making about assessment. This chapter of the *Assessment Handbook* discusses the groups that might be involved and their role in the cycle of planning, implementing, maintaining and evaluating assessment programs.

To the extent possible, allow those who are affected by assessment outcomes (administrators at various levels, teachers, students, parents and other community representatives) to participate in the assessment's design. Each of these groups offers a unique and vital perspective regarding which skills are most important to assess, how they can best be assessed, and how results should be used. Their insights can help ensure that an assessment program is relevant to the needs of the community it serves.

There are a number of issues which could be potential topics of discussion among assessment decision makers. These include:

1. What are the most important skills to assess?
2. What is the most appropriate method for testing selected skill areas?
3. How closely do proposed tests match the curriculum? Is there any way this match could be improved?
4. Which standardized test is the best choice?
5. When should tests be given (both time of year and frequency)?
6. How should results be reported? What data do various groups want and need?
7. How should assessment results be used? Are instructional improvements warranted?
8. What do results indicate about student performance?

9. What improvements in test design, administration, scheduling and so on could be made next time around?

10. How could the assessment program better serve its users?

Not all groups should participate in discussions about all of these questions. Different groups contribute different strengths, and these strengths should be considered when determining how to involve the various constituent groups. The chart on page 4-A shows how the different responsibilities might be distributed across constituent groups.

Committees Serve a Variety of Needs

Depending on a district's particular needs, any or all of several types of advisory groups may be asked to participate in long-range assessment planning. These could include a policy committee, a content/test development panel, and an interpretive panel. Let's consider the composition and function of each one.

Policy Committee

This committee can form an important link between the public and the administrators who guide the assessment. Committee members can interact with community representatives at open meetings or occasionally through surveys, then use their understanding of community priorities to help identify local objectives, draft policy statements affecting the assessment, or review options for testing or remediation.

Depending on district size, a policy committee might comprise anywhere from half a dozen to 25 members or more. More significant than size is the importance of reflecting the ethnic, socioeconomic, sex and age composition of the community. If a district is homogeneous enough, the policy committee can get by with fewer members; as diversity increases, the need for a larger committee increases.

In establishing a policy committee, administrators must remember that many of the persons who serve on such a committee may have limited experience with educational decision making and may not understand all the long-range implications of particular decisions. Their efforts and discussions must be guided, therefore, by someone knowledgeable about assessment methods, administration and planning. The proper function of a policy committee is to make recommendations based on awareness of public priorities and thorough evaluation of the issues involved; final decisions may still rest with administrators and other planners.

Content Panel/ Test Development Committee

The primary function of this group is either to produce tests (if they are being developed locally) or to review tests (if standardized tests are being purchased from a publisher). The composition of the content panel must differ a little from that of the policy committee because more technical expertise is required to actually produce or review tests.

It will be necessary to include persons with subject matter expertise (for example, reading teachers to construct or review reading tests), knowledge of the local curriculum, and experience and skill in item writing (if tests are being developed locally). In addition, it is wise to have persons familiar with the grade levels to be tested and with the relevant student populations. In order to be functional, a content panel should be kept relatively small (five to ten members).

Interpretive Panel

Once testing is completed and results are available, reporting is greatly enhanced if an interpretive panel can be convened to review the results and offer suggestions regarding what those results tell us about student performance. An interpretive panel can bring to light the numerous complex factors affecting student performance and help audiences understand the true relationship among curriculum, instruction and testing.

Another way an interpretive panel might help is by making informed judgments about how well the district's students are likely to perform on each test item. These judgments can be averaged across interpretive panelists and summed to produce estimated total scores. Those estimates can then be compared with actual results to help determine areas of student strength or weakness.

Like the test selection/development committee, the interpretive panel must comprise persons with considerable technical knowledge and expertise. They must, without exception, have thorough knowledge of the content covered by the test. In addition, they must be familiar with the local curriculum and the general capabilities of the students tested in order to have realistic expectations about student performance at each relevant grade level.

Again, it is wise to hold the membership of this group to a minimum (five to ten people) if face to face meetings are anticipated. The interpretive panel deliberations can be undertaken by mail, however, in which case a larger (and therefore probably more representative) group can be asked to participate.

Summary

While final decisions about assessment may rest with administrators, they need not make these decisions alone. Shared responsibility increases support for a testing program through participants' involvement. In addition, as a result of their participation, community members serving on committees also return to the community some understanding of the activities of the district and the constraints under which it operates. In summary, acknowledging multiple perspectives almost invariably strengthens the overall quality of an assessment program.

The School Board's Special Role

A well informed school board can be one of the best allies any assessment program can have. Proactive administrators will strive to keep their boards closely involved throughout the lifecycle of an assessment program. Board review of an assessment program on a regularly scheduled basis can help ensure that testing practices remain responsive to the district's changing needs, rather than becoming entrenched in tradition and, therefore, impossible to modify.

There are three times when it seems especially appropriate to seek school board input. These points occur when the assessment program is in its initial planning stages, when it has been implemented and is operating on an ongoing basis, and when it is ready for review and recycling.

At each stage, the skillful administrator will present the board with information that is relevant to the decisions members must make. In order to heighten the likelihood of the information actually being used, it should be prepared so that it is timely, complete, easily understood, and targeted to the decisions at hand. Let's look more closely at the types of information an administrator might present to a school board during each of the three stages listed above.

During the initial planning stages of an assessment program, the board should be educated about:

- the range of possible purposes for testing,
- the limitations of tests,
- components of an effective assessment program,
- estimated costs for various assessment options, and
- proposed procedures to be followed in test selection or development.

Once an assessment program is in place, the board should be kept informed of its implementation; and when test results are available, the board should receive a report of district performance. This report, delivered prior to public reporting of the scores, should be designed to help members understand the results and to prepare them to receive calls from the public. Such a report might include the following:

- the nature of the district's curriculum and objectives in the areas tested
- a description of the district's teachers, students, and programs
- a summary of test results
- an evaluation of the instructional strengths and weaknesses indicated by the test scores, with possible explanations where discernible
- a display of other educational outcomes, such as curriculum coverage in nontested as well as tested areas, number of graduates, number of dropouts, and so forth
- recommended approaches for correcting identified weaknesses

The assessment program should be reviewed and recycled on a regular basis (perhaps every five years). At that time, various audiences can be surveyed about their perceptions of the program's effectiveness. Questions such as the following might be asked:

1. Do you understand the assessment's purposes?
2. Is the assessment providing you with useful information?
3. What should be added/deleted/modified in order to improve the program?

The results of such a survey could be presented to the board, along with staff recommendations for changes in the program. If the board members have been provided with appropriate information throughout the life of the assessment program, the decisions made at this point should be especially sound ones.

PRACTICAL TIP

Why does testing generate so much controversy? After all, testing is only a major concern for a day or two a year in most districts. One important reason is that people feel that they weren't part of the process, that decisions were made behind their backs, leaving them with only a "take it or leave it" set of test scores. So not only will your testing program be better when all parties are involved from the beginning, but the amount of criticism will be reduced.

Face-to-Face Meetings: Are They Essential?

At least half of Alaska's school districts suffer from having a small number of people spread out over large distances. This makes it very difficult (not to mention expensive) to organize committees that must meet together in person. But the advantages of involving constituent groups in all phases of an assessment program are apparent. Are there ways to involve these groups that don't require face-to-face meetings?

The discussion of interpretive panels elsewhere in this chapter, and in Chapter 10 of the Handbook as well, implies how some committee work can be handled by mail. In fact, independent work by the interpretive panel members may be preferable to having the panel meet together because of the independent judgments that are required.

This idea can be adapted to get feedback on a variety of issues related to assessment programs. In general, it is much less time consuming for group members and much more efficient for the overall process to have people review draft materials rather than to create them.

Review forms should contain questions which are worded specifically to gather the kind of information you need to have; whenever possible, present a set of response choices for each question. Open-ended questions, where respondents simply write their thoughts,

continued, over

are less helpful. In the first place, many people won't take the time to write comments. In the second place, those who do write may or may not address the issues of most concern to you. It's a good idea, therefore, to get the information you must have from closed-ended questions (multiple choice, for example) while allowing people to write in their own comments in addition, if they desire.

Audioconferencing is a strategy used more in Alaska schools than perhaps anywhere else in education. It can't be used for everything, but it is amenable to many types of communication needs. Guidelines for use are available from the LearnAlaska network. As with questionnaires, audioconferencing is perhaps more effective for reviewing suggested content than creating it.

If face-to-face meetings are absolutely necessary, consider arranging a schedule where individuals meet for longer periods of time over a smaller number of meetings. If standardized tests are being reviewed, for example, it would make more sense to bring in representatives from the district's schools for a week during which time they review everything rather than to have people meet one or two days a month over a period of three months.

There's no denying that districts spread over large geographical areas face real problems in involving constituent groups in assessment planning, implementation and review. But in almost all cases, the benefits of involving those groups can be obtained even if people don't actually meet face to face. There are alternatives to group meetings; with the increasing squeezing of district budgets, failing to consider those alternatives cannot be justified.

References

Bollenbacher, Joan K. (no date). Selecting and Using an Advisory Committee, Memo No. 2 in Guide For School Testing Programs, National Council on Measurement in Education, 18-19.

Discusses the "hows" of choosing advisory committee members as well as appropriate ways to use such committees.

Geisler, John (1977, June). A Needs Assessment Approach to Building a School Testing Program. MAMEG Information Report - No. 4, 3 pages.

Presents a checklist that any advisory group could use to determine group consensus as to the test information needs of a school building or system.

Womer, Frank B. (1973). Developing a Large-Scale Assessment Program, Cooperative Accountability Project, 12-22; 25-28.

Discusses how any advisory group can develop questions-to-be-answered prior to development of a testing program. Also stresses the importance of an advisory group.

Participation of Constituent Groups in Assessment Planning

	SELECTION OF EDUCATIONAL GOALS	DETERMINATION OF TESTING PURPOSES	SELECTION OR DEVELOPMENT OF TESTS	IMPLEMENTATION OF TESTING PROGRAM	REPORTING OF TEST RESULTS	EVALUATION OF ASSESSMENT PROGRAM
TEACHERS	Develop goals and objectives	Propose appropriate assessment purposes and content areas to be tested	Review standardized tests/write test items	Administer tests	Interpret scores for students and parents; review instructional progress	Review entire program and suggest improvements
PARENTS/PTA MEMBERS	Rate proposed goals	Rate possible assessment purposes			Review progress of own child, school, district	Review program and suggest improvements
STUDENTS	Rate proposed goals	Rate possible assessment purposes		Take tests	Review own progress; set new goals	Review program and suggest improvements
ADMINISTRATORS	Suggest possible goals	Propose appropriate assessment purposes and content areas	Review standardized tests/oversee test development	Oversee test administration; determine how results will be reported	Review progress of school and district; set new goals	Review program and suggest improvements
COUNSELORS/ CURRICULUM COORDINATORS	Develop goals and objectives	Propose appropriate assessment purposes and content areas	Review standardized tests/oversee test development	Assist with test administration; prepare for score reporting	Make instructional management decisions based on results	Review program and suggest improvements
SCHOOL BOARD MEMBERS	Rate proposed goals	Rate possible assessment purposes and content areas to be tested	Approve test selection	Monitor program	Review district progress	Review program and suggest improvements
COMMUNITY REPRESENTATIVES/ EMPLOYERS	Rate proposed goals	Rate possible assessment purposes			Review district progress	

4-A

District Testing Program Questionnaire

The district's standardized testing program currently is being reviewed. Your opinions about the program and what it ideally should accomplish are an important part of this review process. Please complete the following questionnaire and return it to

_____.

Thank you for your time and your help.

Our district's assessment program ideally should: (Circle one response for each statement.)

1. measure student achievement (i.e., academic strengths and weaknesses). Yes No
2. measure the potential learning ability of students. Yes No
3. provide information that students can use in their choices of specific subjects. Yes No
4. provide information that students can use in making decisions about post-high school education or occupations. Yes No
5. help students understand their own abilities and interests. Yes No
6. help students understand their own achievement levels in various subject areas. Yes No
7. allow students to compare their general level of academic achievement with national norms. Yes No
8. help parents understand their child's abilities and interests. Yes No
9. help parents understand their child's achievement levels in various subject areas. Yes No
10. allow parents to compare their child's general level of academic achievement with national norms. Yes No
11. inform teachers about the abilities and interests of their students. Yes No
12. allow teachers to compare their students' general level of academic achievement with national norms. Yes No
13. identify for teachers possible discrepancies between ability and achievement. Yes No
14. identify for students possible discrepancies between ability and achievement. Yes No
15. identify for parents possible discrepancies between ability and achievement. Yes No
16. allow the district to compare its students' general level of academic achievement with national norms. Yes No
17. provide information that district staff can use in curriculum and program evaluation. Yes No
18. other characteristics:
    _____ Yes No
    _____ Yes No
    _____ Yes No

Our district's assessment program ideally should include: (Circle one response for each statement.)

1. a reading test. Yes No
2. a reading readiness test. Yes No
3. a mathematics test. Yes No
4. a language arts (English) test. Yes No
5. a science test. Yes No
6. a social studies (geography, history, etc.) test. Yes No
7. a writing test. Yes No
8. a "use of sources" test (library skills, graph reading, etc.) Yes No
9. a measure of scholastic ability and/or aptitude. Yes No
10. a measure of student attitudes toward school and learning. Yes No
11. a measure of the self-concept of the student as a learner. Yes No
12. a measure of the student's career interests. Yes No
13. a measure of the student's interest in various curricular areas. Yes No
14. a measure of the student's study habits. Yes No
15. other subject areas:
    _____ Yes No
    _____ Yes No
    _____ Yes No

Please check one category and then fill in the additional information requested.

I am:

___ a student. I am in grade ____.

___ a teacher. I teach _____.
(grade or subject)

___ a parent. My child(ren) are in grade(s)
_____.

___ other (please specify): _____.

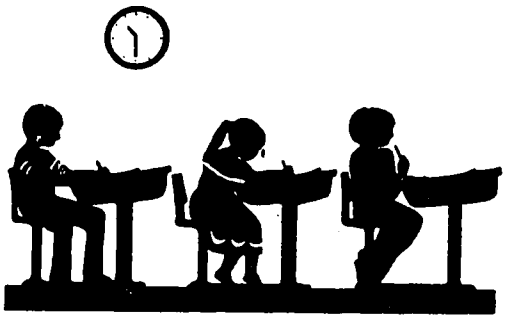
If you have any additional suggestions for the district's testing program, please write your comments on the back of the sheet.

Thank you for your cooperation.

5

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Chapter 5

Standardized NRTs Serve Many Purposes

For most school districts in Alaska--and indeed in the United States--standardized norm referenced tests (NRTs) are an important part of district assessment programs. Such tests are commercially available and have been prepared by measurement experts. Uniform procedures are used in the administration and scoring of the tests, and comparisons of performance can be made because norm groups have been tested using the instrument.

This ability to compare test results means that standardized tests are more useful for some testing purposes than are other types of tests. In any situation where students or student groups must be ranked, a well-chosen standardized NRT has to be given strong consideration as the assessment instrument of choice. As the article on page 3 explains, such situations include selection of students for special programs, guidance decisions, and some accountability and evaluation decisions.

The accountability and evaluation situations where standardized tests are appropriate are those where the question of interest is "How are our students doing compared to similar students throughout the country?" They are much less appropriate when the question is "Are our students learning the skills and obtaining the knowledge we say we are teaching them?"

While there are standardized diagnostic and single-subject tests with associated norms, most districts start with survey batteries which cover a broad range of basic skills content. And many districts add to their standardized testing program an aptitude test, normed on the same sample as the publisher's achievement battery. By so doing, they can compare their students' performance in two ways--externally (compared to similar students nationwide) and internally (compared to their own expected achievement as determined by the aptitude measure). Together, this is a rather efficient use of testing resources for gathering a maximum amount of information.

To get the maximum benefit from a standardized achievement test, the test must measure content that is being taught in the district. Articles on pages 2 and 4, and checklists on pages 5-A and 5-B, provide guidance in selecting the most appropriate standardized test for a district.

Criteria to Use in Selecting Standardized Tests

Recently, a local school district began a review of commercially available standardized tests to see which would be the best choice for their needs. The series then being used in the district had ten-year old norms, which were viewed by district staff as out of date. The committee charged with the review task came up with these criteria for any new test series:

- High content validity
- Co-normed measure of scholastic aptitude
- Norms for individual items
- Recently normed
- High reliability
- Maximum amount of information in minimum testing time
- Individual report forms

Each of these criteria can be thought of as meeting either an alignment, technical or practical need. They form the basis of the Rating Sheet for Standardized Tests shown on page 5-B. Have that sheet before you as we review each category and provide an explanation of the "why's" for some of the entries that might not be self-evident.

Alignment Issues

All of the questions in this section of the Rating Sheet can be answered by looking at the test items; no other manuals need to be referenced. The article on page 4 gives guidance on the preparation necessary.

A critical question--and one that should be answered "yes" if the test series is to receive further consideration -is whether at least 50% of the test items match the district's curriculum objectives. If not, it would be a very inefficient instrument to use; it would give you the comparative information you desire from an NRT, but it would be of little help in curriculum evaluation.

Next, look at how many of your objectives are measured by test items. It is not unusual for a relatively small percentage of objectives to be measured; if the percentage is extremely small for all the test series you consider, though, you might question whether your objectives are too specific and should be broadened.

In some cases, you might be able to use NRT results for diagnostic decisions. To do this, there should be at least three test items per objective (and the more, the better). With fewer than three items, there is too much chance that a student could guess the correct answers and appear to have "mastered" the objective.

The final alignment consideration is whether there is comparable emphasis between test items and district objectives. If you use the form on page 5-A to determine content validity, numbers of your more important objectives should appear in the appropriate column more often than numbers of your less important objectives.

Technical Issues

Good reliability--this is, consistency of test scores--is a necessary, but certainly not sufficient, characteristic of a test. How much is "good"? You should be very skeptical of a test which doesn't have a reliability of .85 or higher (the theoretical limit is 1.0); .90 or better is commonly achieved by national norm-referenced basic skills tests.

The higher the reliability, the more confidence you can have in your test scores. This suggests that you'll need high reliability in those tests which are used for making important decisions. This is especially true when you report individual student scores rather than group scores. Unfortunately, you may have to accept more modest reliabilities when you try to measure more subjective topics--career interest, vocational aptitude, and the like; they don't lend themselves as well to accurate measurement.

Incidentally, subtest reliabilities are always lower than total test reliabilities--reliability depends a great deal on the number of items included in the score--so don't expect every subtest reliability to be in the .90 range. Still, don't let a test breeze through that has lots of low subtest reliabilities or, worse yet, doesn't have any reliability information at all.

Expert opinions about standardized tests can be reviewed in a number of sources (see the Halpern article referenced on page 4). The reviews don't negate the need for a district to study the technical manuals for each series, but they can save a great deal of time in narrowing the field of potential tests.

There is, unfortunately, no standardized test that has a really representative norming sample for rural districts with very small school populations and a high percentage of Alaska Native students. But there are still differences among test series in how the norming groups are put together, and that information must be reviewed. For example, tests which are overwhelmingly normed on students who are mostly from large urban districts would not be as good a choice as tests which have a substantial representation of students from rural districts in the norming sample.

Practical Issues

The practical considerations--including questions about test format, test availability, administration, scoring, costs and publisher's services--may make the difference between an acceptable test series and an unacceptable one. Careful reading of the publisher's manuals (including the administration manual), trying out the test in a real-life application, and talking with other districts using the same test are methods to use in answering questions about practicality.

Advantages and Disadvantages of Commercial Tests

Because of the tremendous resources that test publishing companies devote to the preparation of their standardized instruments, commercially available tests offer several distinct advantages over locally developed tests. But even the high technical quality of these tests cannot overcome certain disadvantages they possess. The list below lays out some of these important advantages and disadvantages.

Advantages of Standardized Tests

- High technical quality
- Content representative of what is being taught in classrooms around the country
- Norms allow comparisons with external groups
- Free consulting on testing issues from publisher's representatives
- No developmental costs

Disadvantages of Standardized Tests

- No reason to expect good match between what is tested and what your district emphasizes in its teaching
- Usually too few items per objective to allow test to be used for diagnostic, certification or certain accountability and evaluation decisions
- May be prohibitively expensive--or even impossible--to get exactly the kind of reports you want
- Recurring annual costs for materials and (perhaps) scoring

While there is probably no way to overcome the last disadvantage, the impact of the first three disadvantages can be lessened by careful planning. Other articles in this chapter tell how.

Consider the eight purposes for assessment discussed in Chapter 2 of the Assessment Handbook. Standardized NRTs are very appropriate instruments to use for some of the purposes, but they are less appropriate for other purposes as can be seen in the following listing.

Usefulness of NRTs

for Instructional Management Decisions

- Diagnosis--LOW
- Placement--LOW
- Guidance--HIGH
- Screening/Admission--HIGH
- Certification--VARIES

for Programmatic Decisions

- Accountability--VARIES
- Planning--LOW
- Evaluation--VARIES

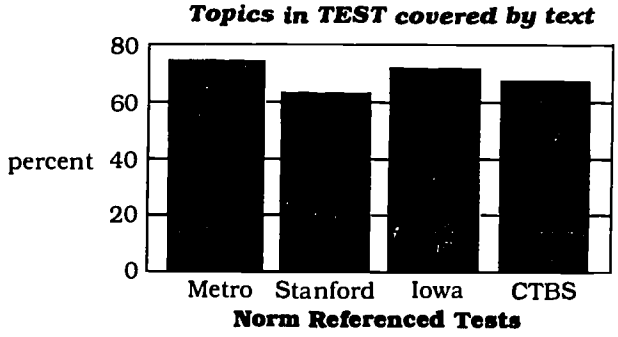
When "VARIES" appears, the usefulness of the test depends on whether decisions will be based on a standard of performance relative to others or on attainment of certain criteria.

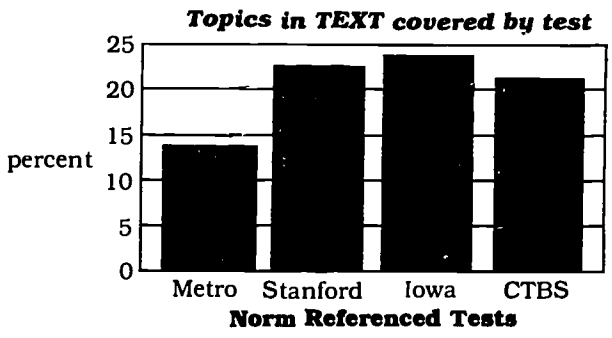
Choosing Among Tests: Does It Really Matter?

There are differences in content emphasis among the various standardized achievement tests. To emphasize this point graphically, look at the two charts to the right. They are based on data from a study conducted by the Institute for Research on Teaching (IRT) at Michigan State University. In that study, the content coverage of commonly used fourth grade math textbooks and standardized tests was compared.

Other texts were included in the study, but we selected Scott-Foresman's Mathematics Around Us for illustrative purposes. By looking at the graph on the top, you can see that for even the best-matched test (the Metropolitan), 25% of the test items are not covered in the text. And the test with the best match on coverage (the Iowa--see bottom graph) still measures less than 25% of the topics covered in the textbook.

This example helps show that no standardized test is a perfect measure of your curriculum; there will always be tradeoffs. But the matching activity described in the article on page 4 is a necessary step in determining what those tradeoffs will be.





MATCHING TEST CONTENT TO CURRICULUM

Chapter 1 of the Assessment Handbook was devoted to the topic of alignment: the matching of curriculum, instruction and assessment. While that is a critical step when tests are developed locally, it is more likely to occur smoothly then because it is the curriculum objectives and instructional materials which are, after all, serving as the basis for test development. But when a standardized test is used, a post hoc matching activity must be undertaken to select the commercial test best suited to the district's curriculum.

A Title I Technical Assistance Center (TAC) provided the following guidelines for matching test content to program objectives; we have edited them slightly to match the form appearing on page 5-A.

1. Number your objectives in the subject area under review. Identify each one as more important or less important.
2. Read the test manual sections that describe the development of the test, the content areas included, and the rationale for the types of items selected. Check to see that the general objectives of the test are in line with your curriculum objectives.
3. Do not rely on the test publisher's description or item classification chart. For each level of the test you plan to use, read each test item and decide whether or not it measures one or more of your curriculum objectives. (If you have many levels to review, consider sampling; randomly select 40-50 items at a minimum of three levels.)
4. For each test item, write the objective number it matches and the degree to which it matches it. Determine how many test items measure a curriculum objective. If over half the test items match no objective, the test does not fit the curriculum very well.
5. For each item, also rate its quality and appropriateness for the intended grade.
6. Determine the content emphasis of the test items. Do more test items match your most important objectives?

By going through these steps, you will have determined the content validity of the test you are reviewing. The form on page 5-A was developed to assist you in conducting the reviews; feel free to reproduce it for as many tests, grades and subject areas as you need.

References

Gronlund, Norman E. (1981). Test Selection, Administration, and Use. Chapter 11 in Measurement and Evaluation in Teaching, Macmillan, 275-302.

Two sections of this chapter deal with "Obtaining Information About Published Tests" and "Selection of Appropriate Tests." There is a suggested test evaluation checklist with 20 categories.

Hall, Bruce W. (1985, Spring). Survey of Technical Characteristics of Published Educational Achievement Tests. Educational Measurement: Issues and Practice, 4,6-14.

Reports the results of a study of the adequacy of test manuals in presenting needed technical information about a test. Includes both recommendations to test publishers and to test users (9 specific suggestions).

Halpern, Marilyn, Mitchell, James V. Jr., and Wildemuth, Barbara M. (1982, Winter). What Tests are Available? Where Can I Find Test Reviews? Where Can I Get a Bibliography on Testing? Educational Measurement: Issues and Practice, 1, 20-23.

Short descriptions of the major sources of information about tests: (1) The Test Collection, Educational Testing Service, (2) Buros Institute of Mental Measurements (Yearbooks), and (3) The ERIC Clearinghouse on Tests, Measurement, and Evaluation (ERIC/TM).

Hills, John R. (1981). Finding, Choosing, and Administering Standardized Tests. Chapter 11 in Measurement and Evaluation in the Classroom, Merrill, 209-229.

Contains a section on general principles to be used in choosing a test. Also discusses sources of test reviews.

Westbrook, Bert W. and Mastie, Marjorie M. (1983, Spring). Doing Your Homework: Suggestions for the Evaluation of Tests by Practitioners. Educational Measurement: Issues and Practice, 2, 11-14 and 26.

A practical, nontechnical, discussion of how to approach the task of selecting the "best" test for one's own use, from among competing tests. Also includes a listing of 26 specific points for comparisons between tests.

CONTENT VALIDITY OF STANDARDIZED TESTS

SUBJECT _____ GRADE ___ _____ DATE _____

REVIEWER _____

TEST _____ FORM _____ LEVEL _____

DIRECTIONS: Write in the item numbers in the "Item" column to match the number of items in the test you are reviewing. Use as many sheets as necessary to cover all the items in the test. For each item in the test being reviewed, consider these three statements:

- The item matches a district curriculum objective.
- The item is of high quality.
- The item is at an appropriate level of difficulty for the chosen grade.

Use the following 5-point scale to show how much you agree with each statement.

5 = Strongly Agree
4 = Agree
3 = Neutral
2 = Disagree
1 = Strongly Disagree

In addition, write the district's objective code that matches the item in the second column of the form.

Item	Curricular Match Obj. No.	Curricular Match How Well?	Item Quality?	OK for Grade?

Item	Curricular Match Obj. No.	Curricular Match How Well?	Item Quality?	OK for Grade?
Totals				

Compare the total scores across tests being reviewed to help make selection decisions.

Rating Sheet for Standardized Tests

Respond to the items below for each test series being considered. (It is not necessary to rate every test level or form.) Write the test series being considered in a column heading, then rate each test using the following scale:

4 = Good
3 = Fair
2 = Weak
1 = Unsatisfactory

(Use the unsatisfactory rating for any missing information, as it would be improper to reward a test that leaves out information. Most publishers of high quality tests know that you need adequate technical information to evaluate their tests; the fact that information is missing reflects negatively on the test.)

The first alignment item must be answered positively for a test to be considered further. Beyond that, the series with the highest rating is most likely the one a district should choose, although individual circumstances may make a district want to weight the criteria differently.

	Name of Test			
ALIGNMENT CONSIDERATIONS				
1. Test items match district objectives?				
2. High percentage of district objectives measured?				
3. Multiple items per objective for [illegible] reporting is to be by objective?				
4. Relative importance of district objectives reflected in test content?				
TECHNICAL CONSIDERATIONS				
5. Acceptable reliability (at least .85 or higher)?				
6. Positive expert opinion?				
7. Normed recently?				
8. Norming sample appropriately representative?				
9. Empirical norm dates match district's testing schedule?				
PRACTICAL CONSIDERATIONS				
10. Format (number of items per page, print size, directions, response mode) appropriate for level of student being tested?				
11. Items free of sex, cultural and ethnic bias?				
12. Easy for teachers to administer and, if necessary, score?				
13. Not too time consuming?				
14. Cost for consumables within budgetary limitations?				
15. Cost for scoring within budgetary limitations?				
16. Score reports that district wants?				
17. Adequate coverage (enough test levels for grades you want to test)?				
18. Test publisher supports out-of-level testing if district wants it?				
19. Alternate forms available if district wants them?				
20. Related tests available if district wants them (e.g., co-normed measure of aptitude or achievement tests for other content areas)?				
21. Consulting and other assistance available from publisher?				
TEST TOTALS				

6

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students

Chapter 6

Assessment Use Tied to Assessment Purpose

We have emphasized throughout this series the importance of identifying the purposes for assessment before even planning an assessment program, much less administering one. Nowhere is the need for this more critical than when it comes to interpreting and using assessment scores. In knowing why assessment is being conducted, answers to questions of test score interpretation become much more obvious.

As you recall, the major distinction between assessment purposes is whether data are being used for instructional management or programmatic decisions. The instructional management decisions--diagnosis, placement, certification, and so on--are made about individual students. Programmatic decisions, on the other hand, are centered on groups of students; evaluation and accountability are served by programmatic uses of assessment data.

Some test scores are appropriate only when they refer to how a single student did on a test; they are inappropriate for describing group results. A later article in this document describes the different types of test scores commonly used in interpreting assessment results. Be sure you are reporting and using the correct type, given your assessment purpose.

And be sure not to overlook the value of reporting trends of test results over time, especially for assessment purposes such as accountability and evaluation. Knowing what progress has been made from year to year is often more valuable than knowing how each single year's results compare to some standard. Ideas for how to display such trends are provided on page 2.

Describing Trends in Test Results

We have said that it is often more important to know how the district's scores are changing from year to year than it is to know each individual year's scores, but what are the appropriate ways to measure such trends? How can you tell if your current students are improving relative to past years, losing ground, or holding their own?

One easy way is to prepare a chart where row headings are grades tested and column headings are

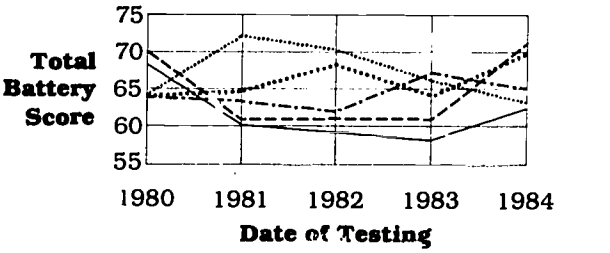
the years of testing (1981, 1982, and so on). Fill in the chart with average percentile scores for each grade each year. (Be sure that the average percentile rank has been computed correctly. It is incorrect to average all students' percentile scores to get the group average percentile; it is also incorrect to average students' raw scores and then refer to individual student norm tables to convert to a group percentile rank. Your best bet is to request group average percentiles when ordering scores and reports from the test publisher.)

To take into account the variation in achievement across students from one year to another, look at the same group of students across years. This is called a cohort analysis. For example, the students who were first graders in 1981 were second graders in 1982, third graders in 1983, and so on. Look to see how the percentile ranking or NCE score of each cohort changes from year to year. Do this for several different cohorts to see if there are trends.

Another way to look at data across time is by preparing a graph such as the one shown below. Plot the scores for each grade for each year's testing (assuming the same test series has been used each year). Over time, you will be able to see if one or two grades are consistently lower than the others.

In the example below, fourth graders' scores are consistently lower than students' scores in other grades. What conclusion should you draw from that information? One conclusion you should NOT draw automatically is that the fourth grade teachers aren't doing their job well. That is, of course, a possibility. But it is much more likely that there is a mismatch between what is being taught in the fourth grade and what is included on the test.

Performance by Grade



Thus, a first action should be to review all of the items in those levels of the test where scores seem abnormally or consistently low. It is likely that some of the items deal with content that doesn't match your curriculum, either because you cover the content at a later grade or the test uses a format unfamiliar to your students. Be aware that if there are many such instances (more than a third of the items, perhaps), you might want to reexamine your curriculum. Test publishers do a pretty good job of reflecting prevalent expert opinion about appropriate curriculum in their tests, and too big a discrepancy may mean that your curriculum is a bit out of step. (That is, of course, your choice. Don't forget, though, that many students don't say in the same district from kindergarten through grade 12. Having too idiosyncratic a curriculum may be a handicap when students transfer to other districts.)

There are other possible reasons for one grade being consistently lower than the others. It might be the first time students use a separate answer sheet or there may be other things about the testing procedures with which students aren't familiar, among other possible causes. In short, don't jump to any conclusions about test performance before looking for alternate explanations.

This is especially important when looking at a graph such as the one first described in this article, where one cohort of students is followed from grade to grade. One cannot ignore the fact that there are differences in student groups. Every teacher has had a group of students that entered the grade better prepared, attended to the lessons better, worked harder and seemed to learn faster that previous classes. But the next year's class might have just the opposite characteristics. So even if the instruction were identical, these two groups would have dramatically different test scores. And the smaller the school or district, the more likely it is that these fluctuations from one student group to another will occur.

So do look at test score trends before deciding just what a single year's test results imply. Look at the same students over time, and the same grades' performance over time. Take into account the different talents of students from one year to another. This added information will make the quality of the interpretations you make about the test scores much better.

Common Mistakes in Working With Test Results

Even if you use the correct score type, there are many pitfalls when it comes time to interpret and use assessment results. What are some things to watch out for?

Perhaps the greatest error comes from making judgments about school experiences by looking at single test scores, such as those collected from an end-of-year standardized norm referenced testing program. Students come to school in many stages of readiness for what they will be taught. What they learn depends not only on what goes on in school but also on what goes on outside of school (the value placed on education by their parents, for example). To judge the quality of a teacher, school or district by relying on just one data point--with no knowledge of the "input conditions"--is analogous to deciding that dairy farmers in Alaska aren't competent if their herds don't

produce the same profit per cow as those in, say, Illinois. Neither judgment takes into account the conditions beyond the control of the person trying to improve the situation. In short, a much more appropriate way to use test scores as a basis for judging educational quality is to look at growth, especially comparing actual growth to expected growth.

A second common error is placing too much emphasis on statistical significance. Statistical significance is a useful concept in guarding against overinterpreting very small differences in scores between groups; determining the statistical significance prevents you from thinking that a small difference that could easily have occurred by chance (because of the unreliability of the testing instrument) is in fact an important difference. The problem is that the determination of statistical significance is so dependent upon the size of the groups being tested. When we compare males' performance on the Alaska Statewide Assessment Test with females', we almost always get statistically significant differences. But that is because there are 3000 or more students in both the male and female groups. Actual differences in the number of questions answered correctly by males and females are extremely small. Similarly, it would be nearly impossible to detect statistically significant differences in males' and females' test scores in Pelican; there are just too few students, even if the actual differences were huge.

The bottom line is that statistical significance is not always valuable. Of much more value is the determination of educational significance, based on a comparison of expected performance to actual performance. Expected performance can be determined statistically or judgmentally. The Interpretive Panel process used in the Statewide Assessment is a good example of a judgmental approach to determining educational significance.

A final common error (final only because our space is limited--not because there aren't other common errors!) is forgetting about the phenomenon called "regression toward the mean." When scores are extreme (the high end or the low end), statistical evidence has shown that such students' scores will be closer to the average score (or mean) if they are tested again on the same instrument. Students with scores at the high end don't necessarily know any less, nor have students with scores at the low end necessarily improved. Be aware of this phenomenon when using test results for any purpose--but especially for program evaluation.

Test Score Types

Six different scores are commonly available from published norm referenced achievement tests. Here is a brief definition of each score type.

The **raw score** is simply the number of questions answered correctly by the student. Raw scores are the basis for all of the other score types; different statistical manipulations are performed to arrive at the other five scores. Raw scores have the advantage that they can be summed over a group of students, then divided by the number of students to get a group average raw score. But it is inappropriate to compare raw scores from one test to another, from one subtest to another, or from one test level to another because each test, subtest and test level has a different number of items. Obviously, a score of 35 is excellent on a 35-item test but relatively poor on most 100-item tests.

Scale scores (sometimes called expanded scale scores, converted scale scores, or standard scores, among other names) express the results from all forms and all levels of a particular test series on one common scale. They are necessary if out-of-level testing is used, and may be desirable even if such testing isn't performed because they allow comparisons to be made from grade to grade or level to level.

A **percentile rank** represents the percentage of students in the norm group who got a raw score equal to or lower than the raw score equivalent to that particular percentile rank. While percentile ranks are easily understood (with proper explanation) by non-educators, they create some problem in that it is not as easy to compute group average percentile ranks as it is to compute averages with some of the other score types. Nevertheless, publishers often provide normative information for percentile ranks for groups as well as for individuals.

A **stanine scale** is composed of nine units. Stanines 2 through 8 are equal intervals; that is, they include the same number of raw score units. Stanines 1 and 9 are larger, though. Some educators like stanines because they are broad enough to prevent overinterpretation of small differences. On the other hand, they are insensitive to small gains.

A **grade equivalent score** represents the middle score of students in a particular grade who were included in the norm group. Grade equivalents suffer from much misinterpretation, especially by parents who think that if their child receives a higher grade equivalent score, the child could be doing schoolwork in that higher grade. In fact, it means that the child got the same score as the older student had the older student taken the lower grade test. Another problem with grade equivalents is that they are inconsistent across grades. A student may score consistently at the 20th percentile at the end of grades 3 through 8. If that student's performance were reported in grade equivalents, however, it would appear that his achievement had steadily declined.

An **NCE (normal curve equivalent)** is a measurement scale developed in conjunction with evaluation requirements for Title I (now Chapter 1). Like the percentile scale, it ranges from 1 to 99; unlike percentiles, though, NCE units are equal in size across the score range. This gives them certain advantages (it is easy to average them across groups, for example) over percentiles. They are particularly useful when you want to aggregate results across tests.

Explaining Test Results to Non-Educators

Most of us have heard horror stories about doing a good job reporting assessment results, only to have those results misinterpreted by someone beyond our control such as a newspaper reporter. What are some things to watch out for--to make sure that they are understood--when providing assessment results to persons who may not have a testing background? Here are several:

1. On a norm referenced test, remember that half the children in the country are "below the norm." The norm is the 50th percentile, the point at which half the scores are above and half are below.
2. Do not compare scores across different tests, or even across different versions of the same test unless conversion tables are provided by the publisher.
3. Your interpretation of results probably assumes that readers know your purposes for assessment. Be sure that anyone else writing about your results also knows those purposes.

Perhaps the most frustrating misinterpretation of test results is when they are used--in isolation--to judge the quality of an educational experience. To use an end-of-year standardized norm referenced test as a primary evaluative tool, without taking into consideration at least the educational attainment of children at the beginning of the year, is naive at best.

It is difficult for district staff to explain this when test results are released, especially if the results are lower than expected or desired; the public thinks the educators are trying to rationalize the poor results. It is a better idea, instead, to have an ongoing public information effort aimed at teaching parents, school board members, media representatives, and any other groups how test scores can and should be used. (For more information on this last type of test score misinterpretation, see Chapter 8: Reporting Assessment Results.)

References

Gardner, Eric F. (1970, January). Interpreting Achievement Profiles--Uses and Warnings. Measurement in Education, 1, 12 pages.

Discusses in detail the use of test score profiles for reporting individual student scores--both uses and pitfalls. Discusses 9 essential points to consider when developing and/or using profile reporting.

Hills, John R. (1983 and 1984). Interpreting Grade-Equivalent Scores (Spring); Percentiles (Summer), Stanines (Fall), IQ's (Spring), SAT/ACT's (Summer), and NCE's (Fall). Educational Measurement: Issues and Practice, 2 and 3, 2 pages each.

A series of six T-F quizzes about how to interpret each different type of test score. The correct answers and an explanation of each answer are provided. Very useful for in-service education sessions. This series will be published as a separate booklet by the National Council on Measurement in Education in 1985 or 1986.

Hopkins, Charles D. (1974). Describing Data Statistically, Merrill, 119 pages.

A short, easily read pamphlet covering an understanding of the statistical concepts (frequency distributions, measures of central tendency and dispersion, test scores, and correlation) commonly used in reporting test scores. Useful both to the neophyte and someone who wants to review previously understood concepts.

Jolly, S. Jean and Gramenz, Gary W. (1984, Fall). Customizing a Norm-Referenced Achievement Test to Achieve Curricular Validity: A Case Study. Educational Measurement: Issues and Practice, 3, 16-18.

Describes how the Palm Beach County, Florida school system matched test questions from the national standardized test they were using to their own local objectives of instruction. The national test was rescored using only items matching Palm Beach objectives, and additional test questions were written to cover local objectives not covered by the national test.

Wilson, Sandra Meacham and Hiscox, Michael D. (1984, Fall). Using Standardized Tests for Assessing Local Learning Objectives, Educational Measurement: Issues and Practice, 3, 19-22.

Describes a specific technique (developed for a Spokane, Washington School District) that shows how local test questions in a standardized test can be matched to a local school curriculum and then scored and reported separately from the total test score.

Interpretation of Test Results: Do You Know the Score?

During the period from Spring 1983 through Fall 1984, a series of quizzes on appropriate test score interpretation was included in *Educational Measurement: Issues and Practice*. An edited sample of questions appear below.

1. Tim is a sixth grader who got a grade equivalent (GE) score of 9.2 on a reading test. This means that Tim could well be put in a class of ninth graders for material in which reading skills were important. T or F?

   False. The GE score cannot be relied on to indicate that ninth grade skills have been mastered. It simply means that if an average ninth grader had taken the sixth grade test in the second month of school, he would have gotten the same number of items correct as Tim did. (It should be noted, though, that often the GEs associated with very high or very low scores are obtained by extrapolation--that is, statistical prediction. It is possible that no ninth grader was ever tested with the test given Tim.)

2. GE scores of 9.2 in reading and 7.3 in math indicate that Tim is farther ahead of his class in reading than in math. T or F?

   False. The standard deviations (the variation or spread among scores) of GE scores vary from one subject to another. The difference between the two GE scores (9.2 and 7.3) may be due to the fact that students tend to differ less within a grade on math than on reading. Because the standard deviations for various subjects differ, we cannot tell whether 9.2 in reading is relatively better than 7.3 in math, and neither necessarily implies that Tim is ahead of his own class.

3. Tim's GE of 9.2 in reading was from fall testing in sixth grade. Tested in the spring, he received a GE score of 8.0. That indicates that his reading skills declined during the school year. T or F?

   False. When GE scores have been extrapolated far above or below a student's grade level, it often occurs that even a single additional item correct can change a student's GE score by more than a year. Tim may simply have gotten one or two fewer items correct in spring than fall.

4. The Jones Elementary School's average GE score in reading in first grade was .6. The average score increased each year until sixth grade, when it was 3.2. Thus the Jones average was .4 year behind at the first grade and nearly 3 years behind by the sixth grade. The Jones students are falling farther behind the national average each year. T or F?

   False. Another peculiar characteristic of GE scores (in addition to the fact that standard deviations vary from subject to subject) is that the standard deviations get larger year by year. Suppose that a person (or a group average) is at and remains at a given percentile score--say the 16th percentile. This same percentile is translated each year into a lower GE score because the standard deviation gets larger from year to year. This can leave the impression that a person (or group) is falling farther behind each year. Similarly, if a student (or group average) is above the mean and stays at the same relative position, he appears to get farther ahead every year in terms of GE scores. This is an illusion created by the GE score system.

5. Susie, a third grade student, scored at the 30th percentile in arithmetic at the end of the school year. Scores this low are regarded as failing, and therefore Susie should be retained for another year in arithmetic. T or F?

   False. Scores at the 30th percentile are really not far below average. Usually no more than a few percent of a class are failed, say 3 or 4 percent, not anywhere near 30 percent. Besides, a nationally standardized test may not accurately sample the arithmetic skills covered in Susie's class.

6. Bill moved from a 90th percentile score to a 99th percentile score from pre- to posttest. Similarly, Jim moved from the 50th percentile to the 59th. They made about equal progress. T or F?

   False. An increase of 9 percentile units at the top or bottom of the scale represents an improvement of many more items answered correctly than the same increase near the middle of the scale. On that basis, one could conclude that Bill made much more progress than Jim.

7. The new principal at Hartford Elementary wants to evaluate the standing of each grade in the school by comparing Hartford students' achievement with the average achievement in a representative sample of elementary schools in the nation. She obtains the percentile scores for each second grade pupil in reading and averages them. The average of these percentiles is the percentile rank for her school's second graders. T or F?

   False. The average of percentile ranks is not itself a percentile rank. To get percentile ranks for averages of percentile ranks, one would have to list the average percentile ranks for all the classrooms in the norming sample and get a new set of percentiles for these average ranks.

8. Mr. Brown learns that the principal wants to compare the performance of each grade in Hartford with other schools. He is correct in claim-

ing that unless the test publisher provides norms on school averages, comparisons of Hartford averages with the average performances in other schools cannot be made. T or F?

True. The scores for a class cannot be averaged and referenced to a norms table for scores of individuals. Class averages can only be evaluated in terms of a norms table for class averages. Some publishers provide such norms, others don't. In the latter case, a comparison between the average score of a class and the average scores of other classes cannot be made.

9. Pedro received a stanine score of 6.5 on a math test. This score should be interpreted as being midway between the sixth and seventh stanines. T or F?

   False. Stanines are represented by single digit whole numbers (1, 2, 3, etc.) and never by numbers with decimal points. Thus, a stanine of 6.5 does not exist. Anyone who uses such a number for a stanine has made an error.

10. A teacher found that most of his students received the same stanine scores in the fifth grade that they got in the fourth grade or even the third grade. He concluded that they are not making much progress in school. T or F?

    False. Tests that use stanine scores refer these scores to students in a particular grade, not to students in general. So a student who regularly receives stanine scores of 5 in a subject from year to year can be assumed to be making normal progress. Normal progress with stanines (or with percentiles or standard scores) is shown by earning the same score over time, not higher scores year by year.

11. Miss Yamini has noticed that the highest scoring third grader who was placed in remedial instruction in reading in her school had a reading comprehension NCE score of 36. She thought the cutoff for remediation was the 25th percentile. She protested to Mr. Wommack, the principal, that an error in placement had been made. Miss Yamini's protest was sound. T or F?

    False. NCE scores and percentile ranks correspond only at 3 points, the 1st, the 50th, and the 99th percentiles (or NCE scores). Actually, an NCE score of 36 corresponds to a percentile rank of 25, which is widely used as a cutoff for remedial instruction.

12. Miss Yamini also collected reading data on her remedial students in the fall of one year and again on the same students in the fall of the next year. She converted her data into NCE scores properly, found the means, and discovered her students had a mean gain of zero. She concluded that her efforts had been in vain because her students had learned nothing. Mr. Wommack comforted her by saying that her students really had improved in reading. He was sure of it. He was right. T or F?

    True. Mr. Wommack is correct. Properly derived NCE scores show a score's relationship to representative national norms. If Miss Yamini used the appropriate norms for the testing dates of her class, zero mean gain in NCE scores signifies that her group improved just as much as the norm group improved. Maintaining the same NCE score does not mean "no growth"; it means "normal growth."

13. Mr. Quigley has noticed that on the Widely Used Achievement Test (WUAT), Sybil's performance on reading has gone up from an NCE score of 10 to an NCE score of 15, and her mathematics performance has gone up from an NCE score of 35 to an NCE score of 40. Mr. Quigley interprets this as an indication that Sybil's scores have increased equal amounts in reading and mathematics. Mr. Wommack says this interpretation is not correct, because it is much easier to improve from 35 to 40 than from 10 to 15. Mr. Wommack is correct. T or F?

    False. Because NCE scores are based on normal distributions, they do not have the property of percentile ranks in which a score difference represents different amounts of change at different places in the distribution. Mr. Wommack would be correct if he were discussing percentile ranks, but not for NCE scores.

14. Mr. Quigley decided to calculate a class mean NCE score by adding up the individual NCE scores and dividing by the number of students in the class. Mr. Wommack was horrified. He said you can't average NCE scores. They don't mean the same thing in different parts of the score scale. Mr. Wommack is correct to be horrified. T or F?

    False. NCE scores are like the other standard scores (z, T, and stanine) in that they are on an equal interval scale. Because they are based on the normal curve (unlike percentile ranks), we feel comfortable in averaging them.

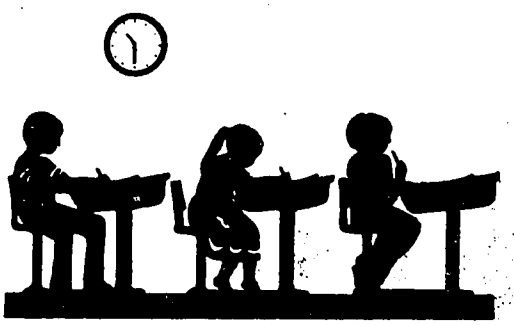
The four articles from which these questions come include a total of 38 questions. Other articles cover SAT, ACT and IQ scores. For information about a subscription to Educational Measurement: Issues and Practices, contact:

The National Council for
Measurement in Education
1230 17th Street, N.W.
Washington, D.C. 20036.

7

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Chapter 7

What Do Testing Programs Cost?

Other chapters of the Assessment Handbook have discussed in broad terms some of the major factors that influence assessment costs: scheduling, frequency of administration, identification of testing populations, and so on. In this chapter, we will attach figures to those factors. The figures that appear in this chapter are approximations, based on average costs for 1985. And in many cases, we have provided time estimates rather than dollar estimates so that each district can put in its own costs.

As much as anything, we want to provide administrators with the information they need to compare relative costs among various components of an assessment program. While it is important to know what standardized tests cost for a district, it is equally important to know that budgeting for the tests is a small part of the total assessment budget--that funds are needed for planning and reviewing the assessment program, interpreting assessment results, writing assessment reports, and so forth.

In general, administrators should plan for the following costs associated with an ongoing assessment program:

- Planning & Test Development/Selection
- Test Administration
- Data Analysis
- Report Preparation
- Assessment Review and Planning

Readers will note that planning appears twice in the list. That is because assessment programs, like most other educational endeavors, need to be thought of in cyclical terms--from planning to implementing to maintaining to reviewing, with the reviewing from one phase serving as the planning for the following phase.

Keeping Costs Low

In a cost conscious society, it is easy to give the cost of testing more weight than it deserves. On a per student basis, testing is a relatively inexpensive educational activity. Further, without testing, the data needed to make sound educational decisions simply could not be obtained.

The point is, cost should not be the major factor in decisions about test selection or development. To select one test over another, for instance, strictly on the basis of cost can have serious consequences. Does a test cover content emphasized in the curriculum? Are the items sound? Does the test measure what it purports to measure? And does it do so in a consistent fashion? Once these questions are answered satisfactorily, cost can be applied as a criterion to select from among comparable testing alternatives.

At the same time, there are certain guidelines which, appropriately applied, can keep testing costs to a minimum. Here are several suggestions:

1. Consider sampling--of either grades or students, or both--whenever testing of every student is unnecessary.
2. Use objective, machine scorable tests whenever possible (that is, whenever direct assessment of performance would provide little additional information).
3. Use separate answer sheets to make test booklets reusable.
4. Consider use of item banks for local test construction.
5. Schedule testing every other year, or according to whatever plan is most economically efficient for the district. Remember that every-year testing is not necessary for all assessment purposes (see Chapter 2: Planning an Assessment Program).
6. Use test data for multiple purposes when appropriate (see Chapter 2).
7. Hire consultants judiciously. In some cases, the cost of a consultant will more than be made up by the time saved. But in other cases, local personnel may prove equally capable.
8. Thoroughly and carefully define the purpose(s) for any assessment in advance. Among the most significant contributors to increased testing costs are the time and resources wasted in administering and scoring tests that were not needed, or that could never provide the data required for the decision at hand.

PRACTICAL TIP

The term "jack of all trades" is particularly relevant to staff in Alaska districts. For this reason, no one can expect district staff to know everything about testing, and the use of a consultant may be in order. But don't hire any consultant without deciding whether their job is to perform a task or to teach district staff how to do the task themselves. It's easiest to have a consultant do all of the work, but that won't increase your understanding of the issues and makes you dependent on the consultant every time the task comes up. It may be more appropriate to hire the consultant to work with several district staff on the project, with the idea that district staff need to perform all aspects of the task themselves in subsequent years.

Local Test Development: An Alternative to NRTs

The focus of this Assessment Handbook chapter is standardized norm referenced tests. But such tests do not fit every need, and some districts have decided to develop their own tests to fill existing holes.

Local test development requires substantial technical knowledge that is beyond the scope of this series to provide. But because locally developed tests do have a place in a comprehensive districtwide testing program, many administrators will eventually need to know something about the topic. The references at the bottom of this article provide some good guidance.

In the opinion of the chapter authors, extensive local test development is not a cost effective activity for most districts in Alaska. At the very least, it is something that should be entered into only when district staff are convinced that they are making the best possible use of their standardized test data. Based on what we have seen (throughout the U.S., not just in Alaska), most districts still have a substantial way to go in tapping the information already available from their NRTs.

- Berk, Ronald A. (Ed.) A Guide to Criterion-Referenced Test Construction. Johns Hopkins University Press, 1985.
- Hills, John R. Chapters 2, 3, 4, and 5 of *Measurement and Evaluation in the Classroom.* Merrill, 1981.
- Popham, W. James. *Criterion-Referenced Measurement.* Prentice-Hall, 1978.
- Rahmlow, Harold F. and Woodley, Kathryn K. *Objectives-based Testing.* Educational Technology Publications, 1979.

Costs of Local Test Development

It is beyond the scope of the Assessment Handbook to provide step-by-step guidance on how a district would develop their own assessment instruments (though some of the references listed on page 4 will be helpful for those who are interested). But it is important for district administrators to have an idea of the scope of the effort they would be undertaking should they decide to develop such tests.

To give readers an example of the steps involved, we will review the process used to develop the Alaska Statewide Assessment Tests. The fact that these tests are used on a statewide basis means that the coverage of review panels and analysis strategies is broader than would be necessary for a district level test. But the steps involved are much the same.

If the number is not preceded by an asterisk, the task can be accomplished by a small group of people (from one to five). The four tasks preceded by an asterisk are better performed by a larger group of district educators or, in some cases, performed by a few people who then submit draft products to a larger group for

review. Please note that the steps listed below do not necessarily happen in sequential order; developing administration directions (step 11), for example, must happen before the tests are piloted (step 7).

*1. Develop/select objectives to be tested. It is important that objectives be broad enough so that several test items can be developed from them, but generally not so broad that a nearly infinite number could be developed.

2. Decide what grade(s) should be tested and how much time will be devoted to testing. Must the entire test take place within two class periods, for example?

3. Develop a content blueprint for each test. The blueprint lists the objectives to be covered and the number of items to be included for each objective.

*4. Develop item specifications. The specifications describe what the item will be like (four-option multiple choice, for example) and provide information that item writers will need to make all the items that measure one objective comparable.

*5. Write or select test items. It has been the experience of some test developers that teachers make better item reviewers than item writers. For that reason, it might make sense to have teachers either select appropriate items from item banks or review items that a person trained in item writing has written, rather than to organize a committee of teachers to write items from scratch. Regardless of how it's done, at least two times as many items should be developed as will be needed for the final test.

6. Format the items into visually appealing tests. In most cases, this means typesetting the items and accompanying them with professionally-drawn illustrations. You might consider formatting the items into two forms (since, based on the recommendation in step 5 you should have at least twice as many items as you need) so that you can choose the best items for the final test. That decision will be based on results of the item analysis (step 8), combined with a reanalysis of the content blueprint.

7. Pilot test the instruments. If test security is an issue, the tests should be piloted in another district whose students are similar to your own. If security is not an issue, the tests can be piloted in your own district. If at all possible, test about 100 students with each pilot version of the test.

8. Conduct item and test analyses. Make sure that the items are performing well from a technical point of view (discriminating between students who really know and don't know the content, for example) and that the test as a whole is a reliable instrument. This task will require access to a computer. (For more information on this issue, see the references listed on page 4.)

9. Select/revise items to include in final test. Based on the item analysis, you may see that some items are fine. But others may need to be modified to solve problems that the item analysis shows. If the modifications are substantial, further pilot testing will probably be necessary.

10. Format tests into final camera-ready copies and print sufficient numbers for the district's testing needs.

11. Develop administration directions and separate answer sheets, if they will be used. If the district will be using a scanner to score the answer sheets, recognize that at least two months should be allowed for printing answer sheets if standard answer sheets cannot be used.

PRACTICAL TIP

The idea of scoring tests with a small optical scanner is a good one, but some caveats are in order. For all practical purposes, you have to use one of the stock forms the scanner manufacturer supplies; it is quite expensive to develop a customized form. Make a review of the available forms part of your purchase decision. And be sure to find out what they cost; if you test frequently, it could still add up to a sizable sum. The most important criterion in your decision, however, should be the test scoring software. It would be expensive to develop your own software; even if you could, it might not work as well as you would like.

---

12. Administer the tests districtwide and score them. With an IBM AT and a Scan-Tron scanner with automatic feed, it was possible to score, tabulate and print reports for, on average, 500 Statewide Assessment answer sheets an hour. Without the automatic feed, that number would have dropped to about 300 sheets per hour. This does not include the time necessary to edit answer sheets (erase stray marks, clean up poor erasures, and so on) before they are put through the scanner, a task which often takes longer than the scanning itself.

*13. Interpret results and produce reports. See Chapters 6 and 8 in the Assessment Handbook for information regarding these activities.

This is clearly a much more involved task than selecting a standardized achievement series and deciding which reports the publisher should provide. At a minimum, a district should allow one school year to get from step 1 through step 10--and that assuming that objectives already exist in the district's curriculum and it is simply a matter of selecting which ones should be included in the test.

Undertaking development in more than one or two subject areas in a given year is not advisable. There are simply too many activities that require teacher and administrator review for multiple tests to be developed at the same time.

References

Gronlund, Norman E. (1981). Constructing Classroom Tests. Part two in Measurement and Evaluation in Teaching, Macmillan, 123-271.

Includes six chapters on test development, with particular emphasis on test item construction (4 chapters). Geared toward classroom application.

Hambleton, Ronald K. (1984, Summer). Using Microcomputers to Develop Tests. Educational Measurement: Issues and Practice, 3, 10-14.

Discusses the use of the microcomputer in item banking and test assembly. Includes a 14-item reference list for further reading. This article is followed by one that critiques the author's major points.

Hills, John R. (1981). Planning the Written Test; Preparing the Test Items; Preparing, Administering, and Scoring the Test; Analyzing, Evaluating, and Improving Tests. Chapters 2, 3, 4, and 5 in Measurement and Evaluation in the Classroom, Merrill, 14-94.

A detailed practical discussion of how one should carry out a local test development project. Following recommended procedures would require a strong commitment to test development.

No authors (1985) Test Resource Catalog. The Riverside Publishing Company, Chicago. Tests and Services for Evaluation. The Psychological Corporation, Cleveland. The Testing Company. CTB/McGraw Hill, Monterey, CA.

Catalogs of three large test publishers. The best source for current prices of standardized tests. Information about other published tests can be secured from the Buros Institute or the Test Collection at Educational Testing Service (see references for Chapter 5).

Womer, Frank B. (1973) Development Tasks in Assessment. Chapter 4 in Developing A Large-Scale Assessment Project. Cooperative Accountability Project, 54-87.

Covers general principles and administrative procedures to follow in developing a large-scale program, but most of the principles also apply to school districts.

Costs Associated with Standardized Tests

Per Pupil Costs for Test Battery (Based on 1985 prices)

TEST NAME	BOOKLETS	ANSWER SHEETS	SCORING	SELECTED SPECIAL REPORTS	COMMENTS
MAT 6 (Metropolitan Achievement Test) Psychological Corporation	$1.08-$1.29 (hand scorable or reusable) $1.54-$1.77 (machine scorable)	$ .27 (hand or machine scorable)	$1.25-$2.15	Group item analysis--$ .50 Frequency distribution--$ .25 Pupil profile--$ .39 Diskette--$ .18/pupil + $6.00 for disk	Separate math and reading tests available for purchase
ITBS (Iowa Test of Basic Skills) Riverside Publishing Co.	$ .87-$1.34 (hand scorable) $ .90-$2.30 (machine scorable)	$ .24-$ .32 (machine scorable)	$ .74-$2.61	Group item analysis--$ .37 Frequency distribution--$ .25 Pupil profile--$ .33 Magnetic tape--$ .14/pupil + $58.00 for tape	Both MRC and NCS scorable booklets available
TAP (Tests of Achievement and Proficiency) Riverside Publishing Co.	$1.34 (hand or MRC scorable)	$ .24 (machine scorable)	$ .74-$1.85	Group item analysis--$ .37 Frequency distribution--$ .25 Pupil profile--$ .20 Magnetic tape--$ .14/pupil + $58.00 for tape	Only MRC scoring available
CAT (California Achievement Test) CTB/McGraw-Hill	$ .81-$ .90 (hand scorable) $1.20-$1.25 (reusable) $1.24-$1.35 (machine scorable)	$ .24 (hand or machine scorable)	$ .75-$1.34	Group item analysis--$ .36-$ .58 Frequency distribution--$ .13-$ .24 Pupil profile--$ .36 Magnetic tape--$ .13/pupil + $55.00 for tape	
CTBS (Comprehensive Test of Basic Skills) CTB/McGraw-Hill	$ .63-$ .90 (hand scorable) $1.13 (reusable) $ .92-$1.35 (machine scorable)	$ .25 (hand scorable) $ .24 (machine scorable)	$ .75-$1.34	Group item analysis--$ .36-$ .58 Frequency distribution--$ .13-$ .24 Pupil profile--$ .36 Magnetic tape--$ .13/pupil + $55.00 for tape	
SRA (SRA Achievement Series) Science Research Associates	$1.21 (hand scorable) $1.17-$1.27 (hand or machine scorable--Levels D-H only) $ .90-$1.24 (with SRA scoring)	$ .40 (hand scorable) $ .56 (NCS scorable; Levels D through H)	$1.00-$1.47	Group item analysis--$ .19-$1.15 Frequency distribution--$ .15 Pupil profile--$ .50 Magnetic tape--$ .09/pupil + $50.00 for tape	
SESAT & TASK (Stanford Achievement Test Series) Psychological Corporation	$1.03-$1.23 (hand scorable/reusable) $1.63-$1.84 (MRC scorable) $1.71-$2.08 (NCS scorable)	$ .22 (hand scorable) $ .27 (MRC scorable) $ .27 (NCS scorable)	$1.25-$2.10	Group item analysis--$ .50-$1.71 Frequency distribution--$ .15 Pupil profile--$ .25 Magnetic tape--$ .15/pupil + $30.00 for tape	

Typical Estimates for Planning and Implementing a District Assessment Program

Developing a comprehensive district testing program can be an involved, time consuming project. Yet to cut corners might make the resultant testing invalid--worse than no testing at all. This chart lists the activities required to develop a sound program, along with suggestions of the staff to be involved in each activity and the typical amount of time required.

Unfortunately, the amount of time required may vary dramatically. In districts where the responsibility is concentrated in a few people, meetings will be brief. And in some Alaska schools, there may be only one person who takes care of all aspects of the testing program. Further, administering tests to 15 students is obviously nowhere near as involved as administering the same test to 15,000. In short, take the estimates given below as a starting point and modify the estimates based on the complexity of the development task for your district.

I. ASSESSMENT PLANNING

- **Develop overall plan for assessment**
  Resources needed: Planning Committee--five to ten members for one or two days
- **Select standardized tests**
  Resources needed: Content Review Committee--five to ten members for two or three days
- **Design score reporting and recordkeeping systems**
  Resources needed: Assessment Coordinator--one or two days
- **Coordinate and document assessment planning activities**
  Resources needed: Assessment Coordinator--two to four days; Support Staff--two to five days

II. TEST ADMINISTRATION

- **Order test materials and score reports**
  Resources needed: Assessment Coordinator--one day
- **Receive test materials and package for distribution to schools**
  Resources needed: Assessment Coordinator--one day; Support Staff--one-half to four days
- **Prepare testing schedule and test administration guidelines**
  Resources needed: Assessment Coordinator--one day; Support Staff--one or two days
- **Train teachers or school site coordinators to administer tests**
  Resources needed--Assessment Coordinator: one-half day per group to be trained
- **Oversee test administration**
  Resources needed: Assessment Coordinator--one-half to three days
- **Collect test materials and prepare answer sheets for scoring**
  Resources needed: Assessment Coordinator--one day; Support staff--one-half to six days

III. DATA COLLECTION/ANALYSIS

- **Machine score tests in district (if applicable)**
  Resources needed: Assessment Coordinator--one day; Support staff--one to ten days
- **Review and interpret test results**
  Resources needed: Interpretive Panels--five to ten members for one-half day
- **Distribute results to schools**
  Resources needed: Assessment Coordinator--one-half to one day

IV. REPORT PREPARATION

- **Write district-level report, detailing results in all areas**
  Resources needed: Assessment Coordinator: two to four days; Support Staff--two or three days
- **Prepare press releases, newsletter articles, etc., summarizing highlights of assessment**
  Resources needed: Assessment Coordinator--one day; Support Staff--one day

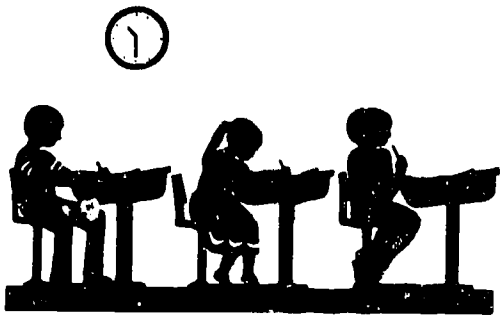
V. REVIEW AND PLANNING

- **Review overall plan for assessment**
  Resources needed: Planning Committee--five to ten members for one to two days
- **Coordinate and document assessment review activities**
  Resources needed: Assessment Coordinator--one to two days; Support Staff--one to two days

8

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Chapter 8

Different Publics Demand Different Reports

We have emphasized throughout this series that the only reason for administering an assessment instrument is to improve decision making by having assessment data. It goes without saying that assessment results must be reported to at least the person(s) who will be making decisions based on the data. And at the district level, reporting usually goes beyond that.

Details included in reports vary according to the needs of the publics for which the reports are intended. The public's characteristics--their responsibilities, time available for learning assessment results, prior knowledge, and numerous other variables--should be considered when planning reporting strategies.

Ask yourself four questions when you first start to plan the release of assessment information:

- Who needs to know this?
- When should they first hear about it?
- What is the best way to get this information to them?
- What will they be most interested in?

The following ten publics may be answers to the first question. In the paragraphs below, items that might interest each group are provided. Other sections in this chapter provide guidance on how to answer the second and third questions.

Students need to know how they are performing.

Parents need to know how their children are performing, how the educational program as a whole is functioning, and how their children's test scores compare to those of children in other schools or communities.

Teachers need information about the quality of instructional decisions they are making about students.

Principals or others responsible for supervising instruction need data to determine whether programmatic and instructional goals and objectives are being met.

Curriculum specialists need to know whether current goals and objectives are being met (or if they are even the correct ones), whether current materials and methods are appropriate, and where technical assistance should be targeted.

The **superintendent** needs data that will help the school board make basic funding and staffing decisions, and to approve specific instructional or curriculum decisions. The **School Board** needs similar information, though perhaps in less detail.

Special interest groups such as parent advisory committees need to know how well students are performing, their areas of strength and weakness, and how the schools intend to maintain strengths and improve weaknesses.

The **news media** need to know similar information as the special interest groups but it is a good idea to give them background information so they understand testing terminology and what test information can and cannot be used for.

The **community** wants to know how local students are performing, compared perhaps to local expectations and also to student performance in other localities across the state and nation.

For all of these groups, data must be interpreted as they are being reported; no test data are meaningful in their raw state. Chapter 6 of the Handbook, Interpreting and Using Assessment Results, discusses this topic in detail.

Nine Steps To A Positive Press

Ned Hubbell, a school public relations consultant in Michigan, has identified nine steps as necessary to obtaining positive response when releasing test results.

1. Let people know **ahead of time** that you'll be releasing results.
2. **Identify** the publics to whom you'll report.
3. Plan a **timetable** for release of results.
4. Help those who need to **explain results** to others.
5. Help the media **interpret** to the public.
6. Start with **simple explanations**.
7. **Summarize** what the results mean.
8. Tell **what will be done** with results.
9. Take testing results to **targeted groups**.

These steps will be discussed further in other sections in this chapter. Keep these nine steps in mind, and many of the problems which often accompany the release of assessment results will be short-circuited.

Reporting Within the District

Assessment results should be released to school staff members before they are released to the public. This could be done through briefing sessions, regular school publications and specially prepared materials. It is a good idea to have an oral presentation, especially for building principals, so that questions about the results can be raised and answered. Principals can be helpful in briefing their staff members. Prepare graphs and charts that principals can use when they talk with their staff and parent groups.

Remember that staff means every employee in your school district, including janitors, cooks, bus drivers and secretaries as well as certificated personnel. Research shows school staffs are among the most credible sources of information--more credible than the local news media--and further, that classified staff frequently have more credibility with the community than teachers and administrators. So don't write any staff members off as news sources just because they aren't certificated educators.

Reporting to Parents and the Broader School Community

If the first step in a good communications program requires school staff members to understand assessment results, the second step is to be certain that parents understand them. Consider carefully the information that is provided to parents about their children's test scores.

It is highly unlikely that the scores are adequate by themselves. A cover letter accompanying the scores could include simple explanations of the type of test scores reported (percentiles and standard scores, for example), what it means to be above or below the norm, and how the scores are used. A sample of such a letter is provided on page 8-b of this chapter.

School newsletters are another good way to share information about assessment results. There may be more space available for describing results in a locally produced newsletter than there would be in a newspaper or similar public media source. Similarly, if the newsletters are produced by schools rather than by the district as a whole, information more directly relevant to an individual school can be highlighted.

PRACTICAL TIP

Media representatives always want to hear from the highest ranking district administrator possible. Have press conferences--and other meetings related to assessment results--led by the superintendent or assistant superintendent. But be careful that this approach doesn't backfire; be absolutely certain that the presenting administrator understands the test results. No one expects superintendents to be aware of every technical detail of the testing, but they must always understand what the results say about the performance of their district's schools. It is a mistake to try to bluff one's way through a press conference, so if it comes down to a choice of having an uninformed superintendent make the report or an informed lower level person, it's usually best to go with the person who knows what's going on.

Report With Pictures, Not Just Words

Graphs, charts and other visual representations of test results not only add interest to reports, they also help explain those results. And graphs and charts are invaluable when making oral reports of test results because they allow the audience to see as well as hear the results.

The first common error people make when developing charts is to put too much information in one chart. This nearly always results in a busy, unreadable graphic. On the other hand, selecting a subset of data requires the person producing the chart to choose an aspect of the data to illustrate. Deciding which graph to use may be the most difficult decision of all.

The box at the right lists some commonly graphed assessment results and the type of graphic that is usually most appropriate. Examples of some of these types are provided on page 8-a, along with some common errors in their presentation.

- Comparisons among groups at a single point in time (For example: How do district results from the Spring 1984 testing compare to results from the state and nation?)--**bar graph**
- Parts that make up the whole (For example: What portion of all students tested have been in district schools since kindergarten?)--**pie chart**
- Trends or comparison in trends over time (For example: How do achievement gains over time compare for various groups of students?)--**line graph**
- Measures of association (For example: How is attendance related to test scores?)--**scattergram**
- Likelihood that results are "real" and not due to chance (For example: Do the eighth graders' higher scores represent real growth?)--**error band chart**

News Release Pointers

A news release should be written and distributed to newspaper reporters and other media representatives when you are ready to publicly announce your test results. Preparing effective news releases requires attention to both format and content. In this article, we'll look at both.

We've all heard about the five Ws--who, what, when, where and why. Answers to these questions--plus a sixth question, how--is the information that should appear in the first paragraph or two of news releases.

PRACTICAL TIP

Ask a few people who are members of your intended audience to read a draft of your reports before you produce the final version. When you are presenting complicated statistical or evaluative data, these people will help you learn whether what you've said will even be intelligible to a lay audience. Don't assume that because you think a report reads well, it will be readily understandable to the Board, other district staff or the general public.

Each succeeding paragraph should be of declining importance. In this way, editors will not omit vital information if they have to cut the story to meet space limitations. Also, readers can get the gist of a story if they read just the first few paragraphs.

There are several other format guidelines for preparing news releases, as follows:

- Double space your story on one side of the paper only, using school or district letterhead, and leave generous margins on each side of the page to allow for editing.
- The source of information (name, title and telephone number) should be in one corner (under letterhead) of first page.
- End each page with a complete paragraph. If a second page is necessary, always type "MORE" at the bottom of the first page.
- Use short words, short sentences, short paragraphs.
- It's a good idea to include quotes, but be sure to completely identify all persons quoted and spell their names correctly.

When preparing news releases, it is important to briefly describe the purpose of the testing, who was tested, and when the testing occurred. In addition, the practical use of the results should be summarized.

Be sure to describe both positive and negative results. Don't explain the negative results away. Instead describe the district's plans for improving those areas where weaknesses were shown.

When discussing negative results, outline the non-instructional problems the school and community must address. These might include:

- Absenteeism
- Physical well-being of pupils
- Parental support of school instruction
- Pupil interest and motivation
- Pupil mobility

And if explanations can be offered for improved scores (for example, a new reading program or cross-age tutoring in math), mention them too.

To make sure your story is understandable, ask several parents to read it before submitting it to the newspaper. They can often point out confusing sections which you, because you are so close to the data, are unable to see.

References

Badal, Alden and Larsen, Edwin P. (1970, May). On Reporting Test Results to Community Groups. Measurement in Education, 1, 12 pgs.

A detailed discussion, with charts, of how the Oakland, California Public Schools routinely report school-wide test results to the public. Suggest a variety of techniques.

Iverson, Grace (1984, Summer). Raising Test Scores. Educational Measurement: Issues and Practice, 3, 45-46.

Describes a Lansing, Michigan school district plan for helping to improve its Michigan Statewide Assessment scores. It includes a description of working with the local newspaper to publicize the plan, implement it, and report test results.

Lenke, Joanne M. and Beck, Michael D. (1980). The Ways and Means of Test Score Interpretations. In New Directions for Testing and Measurement: Interpreting Test Performances, Jossey-Bass.

Discusses a variety of means for reporting test scores both to school staff (classroom, building, system) and the public (student, parents, general public). Also looks to future trends in this area.

NCME Award (1983, Fall). A Public Dissemination Plan of Test Results and Application of Testing to Improvement of Instruction. Educational Measurement: Issues and Practice, 2, 15-20.

Describes an award-winning plan for disseminating school district test results to the public (using school-printed inserts in a local newspaper). Contains a mailing address for securing a sample copy of "APS in Action" (Albuquerque Public Schools).

Ricks, James H., Jr. (Undated). On Telling Parents About Test Results. Test Service Bulletin No. 54, The Psychological Corporation, 4 pgs.

An old but still timely discussion of how to communicate test scores to parents. Includes specific suggestions for translating test "numbers" into test interpretation "words."

Chart 1 - *Poor*

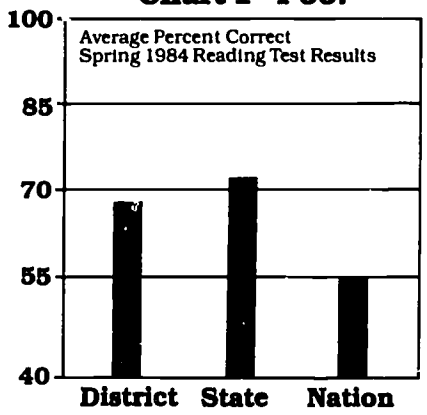
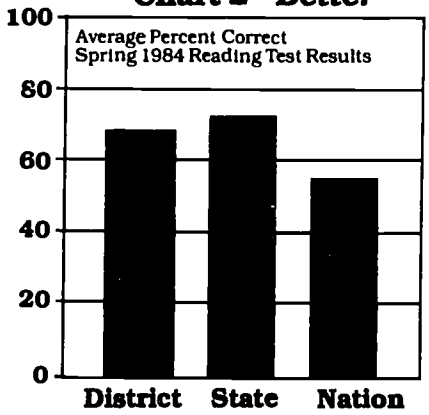


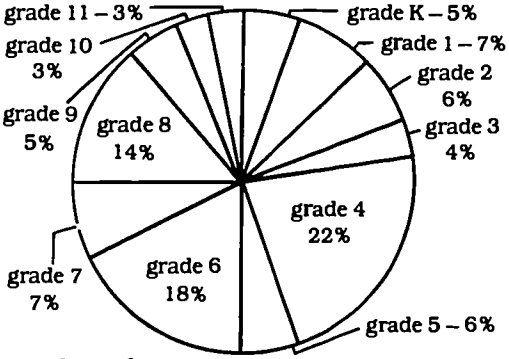
Chart 2 - *Better*



What makes Chart 2 better than Chart 1?

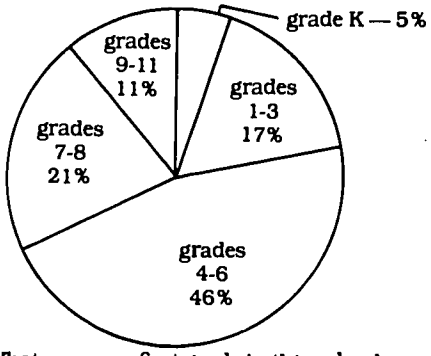
- Bars are wider than the space between them.
- Zero is included on axis. (Zero *can* be omitted, but break in axis should alert reader.)
- Grid lines don't pass through bars.
- Y axis uses scale that makes bars easy to interpret.

Chart 1 - *Poor*



Class of 1985's responses to question regarding their entry into school district

Chart 2 - *Better*



What was your first grade in this school district? (Response from Class of 1985)

What makes Chart 2 better than Chart 1?

- Too many segments in Chart 1 — limit pie charts to five or fewer segments.
- Chart 2's title is easier to understand.
- Labels should go inside segments when there is enough space.

Chart 1 - *Poor*

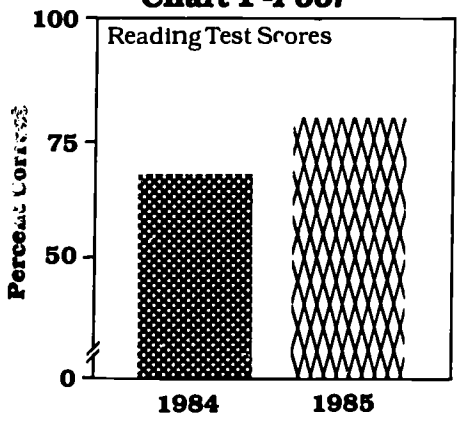
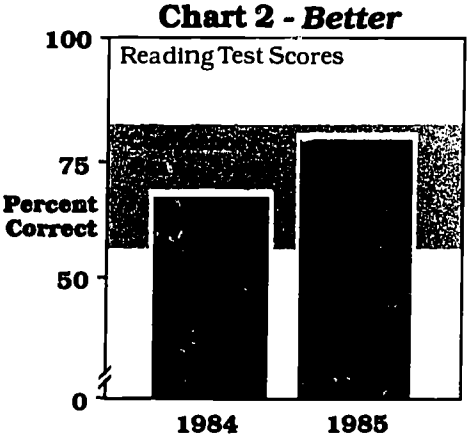


Chart 2 - *Better*



What makes Chart 2 better than Chart 1?

- Confidence band (determined using standard error of measurement) shows that 1985 score is not *really* different from 1984's.
- Y axis is labeled horizontally.
- Patterns used for Chart 2's segments don't distract reader.

Chart 1 - *Poor*

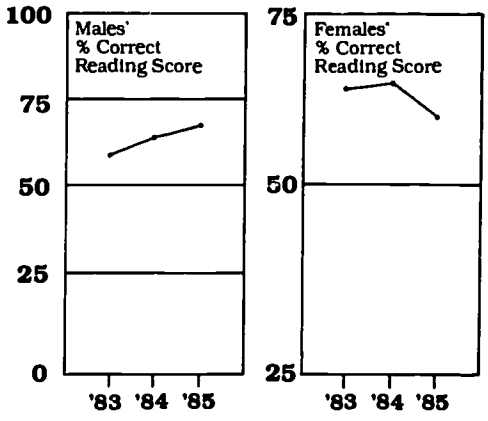
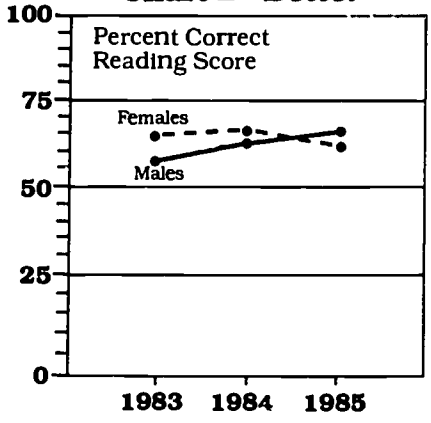


Chart 2 - *Better*



What makes Chart 2 better than Chart 1?

- Males' and females' scores are compared on the same scale (see Y axis).
- Axis numbers are large enough to read easily, and enough tick marks are added to aid interpretation.
- Lines showing data are thicker than grid lines.
- Each data line is labeled.

REPORT SAMPLES

• Here is an example of a letter accompanying test scores that might be sent to parents.

Dear Parents:

The enclosed test information sheet summarizes the results of the achievement test your child was administered last April at [redacted] School. The test covered reading, math and language arts and has been normed on a national population. This means that we can compare how [redacted] students performed on the test to a nationally representative group of students at the same grade levels.

A detailed explanation of the scores is provided on the reverse of the information sheet. Of particular importance is the percentile column, which shows how your child's score compares with others at the same grade in the national norm group. If your child's percentile score is 75, it means that he or she received a score which was higher than 74% of the norm group that took the same test; on the other hand, it means that 25% of the students in the norm group had a higher score than your child's.

While the scores for an individual child may not reflect his or her true ability, I feel the test results are accurate for most children and for the school as a whole. Speaking of the school's average results, we learned that our students are doing quite well -- above the national average -- in most sections of the reading and language arts tests, but math continues to be a problem. [redacted] students seem to be having particular trouble with math application items -- what we call story problems. Understanding the words in the items doesn't seem to be the problem, as our students' reading comprehension scores are among the highest. It just appears that they are having trouble picking out the important information they need to work the math problems embedded in the stories. This is an area that most of our district students -- not just those at [redacted] -- seem to be having trouble with, and it has prompted us to undertake a comprehensive review of our math curriculum and instructional practices. We'll be reporting to you later to tell you what we have decided to do to improve our students' math skills.

Please feel free to contact your child's teacher or me if you would like further help in interpreting your child's scores or want to know anything else about the [redacted] assessment program. We would be pleased to talk with you.

R. J. Page, Principal
[redacted] Elementary School

• A report on test results in a school district newsletter might be worded this way.

SRA test results show achievement above norm

[redacted] students in grades three through nine are tested each spring for their command of basic skills.

The 1983-84 test results again show that [redacted] students achieve well above the national norm; a level of performance the District has come to expect throughout the years.

Referring to the continual improvement in test scores, Ray Brown, [redacted] measurement specialist said, "Basic skill development has continued to be the primary emphasis in instruction at the elementary and intermediate school levels in [redacted]."

The tests, Science Research Associates (SRA) Achievement Series, are primarily designed to measure the extent to which students have acquired skills in reading, language, math, reference materials, science and social studies. Educational Ability Skills (EAS) are also measured.

"[redacted] elementary students, while continuing to maintain high achievement levels in reading and language, have markedly improved in their mathematics skills," Brown added. "This is not merely an improvement in computation skills, but in math concepts and problem solving ability."

Intermediate students continue to excel in reading and math with improvement shown in language mechanics. "While students did not show the desired total amount of improvement in spelling this year, never the less, there has been some improvement in that area since the implementation of the Cedar Rapids Spelling Program," Brown said.

However, in grades 6-9, more than twice as many [redacted] students score in the highest achievement range overall, than is the case nationally.

"Achievement scores are but one indicator of the effectiveness of an educational program," Brown emphasized. "No measuring instrument taken by itself can assess the wide diversity of programs and offerings in the District. However, the scores do indicate that [redacted] students score well above the national norm. Test scores indicate [redacted] students are well rounded in basic skills.

A complete report on the SRA test was made to the School Board earlier this fall.

• A news release on test scores might be presented this way.

[redacted] District Student Test Scores Improving

Student test scores, while still below the national average, have steadily improved over the past four years in the [redacted] School District, special services coordinator Ed Smith has told the district board.

Students in fifth, eighth and tenth grades took California Achievement Tests last March. Overall scores were one to two years behind the national average for all three grades but in almost all cases, improvements over last year's scores were made in both overall scores and various subject areas.

Superintendent Ray Sanders pointed to several things contributing to increased student achievement in [redacted]:

- Clear curriculum goals supported by appropriate instructional materials.
- Improved student discipline.
- Better classroom management.
- More training and updated information for teachers.
- Prompt, accurate and frequent feedback to students, teachers and parents.
- More attention to individual student needs.
- Partnerships with parents and community groups in supporting student learning.

"We have really just finished the curriculum goals work that has kept our staff busy over the last five years. In those areas where goals and supporting instructional materials have been in place for at least three years, we have seen [redacted] students make dramatic achievement gains. I am confident that those gains will continue now that the complete curriculum has been specified and we turn our attention to providing additional resources to the educators who teach that curriculum," Sanders told the school board at last Tuesday's regular meeting.

When board member Elizabeth Sadler asked what parents could do to most help their children achieve higher scores, Sanders replied that getting students to school was the single biggest thing. "Students who are absent from school -- whether on an excused absence or otherwise -- really suffer because it's very difficult to catch up when a lot of days have been missed. We know that research has shown that 'time on task' -- the amount of time students are actively engaged in learning -- is a critical factor in their achievement. If students aren't in school, there's just no way that their time on task can be high."

9

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students

Chapter 9

Why Keep Records?

After the score reports from an assessment have been distributed to parents, students, teachers and administrators, what should happen next? What sorts of assessment records should a district maintain and for how long? What's necessary to keep and what can be thrown away?

It is important to realize that records shouldn't be kept just because "it's always been done." Instead, the various reasons for maintaining records should be determined. Then recordkeeping procedures appropriate to each information need can be established.

Important reasons for keeping assessment records include the following:

- Parents and students expect the schools to keep test scores as part of a student's cumulative record.
- Teachers need information about student achievement in order to make effective instructional management decisions.
- School and district staff need information about pupil progress in order to evaluate the effects of curriculum and instruction.
- State and federal programs may require student achievement information as part of their mandatory evaluation requirements.
- A district may want to monitor the effects of its programs over time, in order to determine how well its schools are doing.

What Types of Information Should Be Kept?

A recordkeeping system needs to be comprehensive enough to maintain all the various types of information that will be required. The assessment component of the system might include information about individual students, including background demographic characteristics and scores from administrations of district and state tests as well as information about student test performance aggregated at the class, school and district level.

Additional information that would be useful when considering the meaning of test scores includes the following:

- Information about each school's instructional programs, including curriculum goals and objec-

tives, textbooks used, instructional approaches adopted, and any special programs in effect.

- Information about changes that have occurred that might be related to student performance, including new curriculum or instructional efforts, new textbooks, changes in teaching staff, changes in instructional time, shifts in tests or testing procedures.

The following three steps should guide the design of a comprehensive recordkeeping system:

1. Determine all of your district's information needs, at individual student, classroom, school site and district level.
2. Verify that any mandatory requirements associated with state or federal programs will be met.
3. Design data gathering and storage procedures that will meet your information needs in the most efficient and cost effective manner.

How Long Should Assessment Records be Kept?

One district successfully manages its assessment data by following these guidelines:

- Score reports are returned to students, parents, teachers and administrators as soon after testing as possible.
- Individual student score reports are entered into each student's cumulative record.
- The district office maintains an extra copy of individual student score reports for one year after the assessment date.
- Grade level score reports (e.g., group performance summaries) are maintained in the district office until the students in that grade have graduated. For example, score reports from this year's first graders will be kept until those students graduate from high school.
- Grade level score reports are stored in notebooks by year (for example, all the score reports from this year's assessment, grades one through twelve, are filed in a single binder).

Using Assessment Records: Three Scenarios

Assessment records, no matter how comprehensive or well maintained, are not, by themselves, of much utility. They only become beneficial when the data they contain are used to answer important educational questions. The following three examples demonstrate some effective uses of assessment information as a part of educational decision making.

In Evaluation

A new mathematics textbook series was adopted two years ago for grades five through eight. Since then, district math scores in those grades have dropped. An evaluation is designed to answer the question, "Was the new textbook a mistake?"

District assessment records will be used to determine students' math test scores for several years before the new text was adopted, as well as student performance in other subject areas. The district records will also provide descriptive information about student background characteristics. This information will be used in conjunction with such information as how well the test measures important district goals, how well the test matches the textbook, and teachers' subjective opinions about what has happened. Although assessment records will not be the sole source of information for this evaluation, they will yield much of importance.

In Curriculum

When one year's statewide assessment scores showed study skills to be a weak area, a district decided to examine its language arts curriculum. Some definite weaknesses were found in study skills, and these were addressed by developing and implementing new objectives. Much to teachers' delight, the next set of assessment results showed a definite improvement in study skills, although isolated areas remained weak.

Now the curriculum committee will meet again to examine the remaining areas of weakness and determine if they warrant more attention. In two years, the newest assessment scores will once again be examined to judge if the district's progress has been adequate. By using assessment records within the context of its own particular objectives and goals, this district has demonstrated another way in which adequate assessment records can help inform educational decisions.

PRACTICAL TIP

Many schools in Alaska have considerable turnover in the teaching staff from year to year. Don't make new staff start from scratch to find out their students' current achievement level. Good assessment records can be extremely valuable to a teacher who doesn't know where the students stand.

In Instruction

Because problem solving is an important focus of a district's mathematics curriculum, a special computer-assisted instructional program is adopted. Equipment is limited, meaning that only about half the students can receive the program. The selection of program participants is greatly facilitated because the school has maintained complete records of student scores in problem solving on the district's standardized test. These records are now consulted in order to determine which students will most benefit from the computerized instruction. Teachers are also asked for their estimates of students' problem solving ability. Used in combination, teacher judgment and assessment records provide a better instructional decision than might have been made using either source alone.

Computers Can Help in Recordkeeping

Recordkeeping requires the maintenance and manipulation of data, tasks at which computers excel. And advances in computer technology mean that $10,000 or less will buy a computer capable of handling the assessment records for even the largest district in Alaska. But certain cautions are in order.

- The nationally published computer-based recordkeeping (CBRK) systems are not always well designed, and even the best of them may not be suitable for a given setting.
- The computers already in schools (including the Apple II) are often not suitable for detailed recordkeeping with many students. A disk-based Apple II can work with about 100 students at a time if each record is a thousand characters or less in length. Be very careful that your records will fit. There are tales of schools that converted to a CBRK system, spent months preparing and entering data, and then found, when entering the last quarter of students, that the program's storage space was full and there was no way to include the rest of the students.
- The system may not be flexible enough. Some systems have limited reports they produce and inquiries they permit. If you want something else, you're out of luck.
- The system may not have any advantage over paper-based systems. Unless the computer manipulates the data to produce new information, a paper-based filing system is just as practical. Make certain the CBRK system does something other than just store information.
- Staff may not be committed to the system's use. If one teacher doesn't want to enter the data, or one set of test scores isn't recorded, the integrity of the entire system is jeopardized. A computer-based system sometimes makes recordkeeping more formal and intensive than staff are willing to undertake. Be sure to find out the level of commitment in advance.

Large Sites versus Small

What's practical in a school or district with a large number of students may not be practical in a smaller site. It would be foolish to implement an extensive computer-based system in a district with only five or ten students in a grade. Under such circumstances, printed reports are easy to use for reference; further, there is not much information to manipulate (for example, averages are easily done with a calculator). And should some information necessary for decision making be missing, it's relatively easy to collect the information from other existing sources.

The converse is also true. A district with thousands of students cannot successfully manipulate assessment records by hand. (Just one year's statewide assessment records for Anchorage, for example, have over half a million pieces of data.) Nor can information be collected on an ad hoc basis. The recordkeeping system for such a district has to be established in advance, in a formal manner and probably with the help of experienced data processing personnel.

Student Records and Compliance with the Law

Administrators need to comply with two federal laws when designing a recordkeeping system. The Hatch Amendment, with its associated 1984 Department of Education regulations, and the Family Educational Rights and Privacy Act of 1974 (the Buckley Amendment) both impact heavily on the handling and release of student records. These acts are designed to give parents more control over the testing and teaching of their children. In addition, they give parents and students access to their educational records and the right to privacy regarding the dissemination of records containing personally identifiable information. The major provisions of each law are summarized below.

The Hatch Amendment

- Parents must be given a chance to inspect instructional materials and give their consent before students take part in a wide range of classroom activities or use materials in programs receiving federal funds.
- Parents must give consent before their children submit to "psychological tests or treatments" in areas that include potentially embarrassing psychological problems, anti-social or self-incriminating behavior, criticisms of family members, and statements of family income.

The Buckley Amendment

- Educational records must be released upon request to parents (including a noncustodial parent) or students 18 years of age or older.
- Personally identifiable information in student records may be disclosed only with written approval of parent.
- Parents and students are allowed to correct errors in students' records.
- School officials with legitimate educational interests are allowed access to educational records of a student without prior parental approval.
- City or state police officers and potential employers are not allowed to have routine access to student records.
- Federal funds can be withdrawn from a district for noncompliance with the regulations.

Suggestions for minimizing compliance problems with the two laws include publicizing parents' and students' rights in school publications (for example, parent or student guides) and establishing a consistent policy for discussion and complaint.

References

Gardner, Eric (1982). Some Aspects of the Use and Misuse of Standardized Aptitude and Achievement Tests, in Ability Testing: Uses, Consequences and Controversies. National Academy Press, 315-332.

A detailed look at common criticisms of standardized tests. Good background for determining what to record from testing programs and how to record it.

Lyman, Howard B. (1980). Metrics Used in Reporting Test Results, in New Directions in Testing and Measurement: Interpreting Test Performance. 17-34.

Discusses in detail the three basic types of scores used to describe test performance. Helpful in deciding how a local district should record its own student scores.

Sampson, James P., Tenhagen, Carl A. and Ryan-Jones, Rebecca (1985). Guide to Microcomputer Software in Testing and Assessment. Special issue of AMECD Newsnotes. 20 (August), 12 pages.

A listing of about 100 software programs for administering, scoring, recording, and profiling results from a variety of tests. Also lists computer software vendors and includes a 266-item bibliography.

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students

Chapter 10

What Role Can Statewide Assessment Play?

Every other year, the Alaska Department of Education tests every fourth and eighth grade student in the state on important reading and math skills. Results are used at the statewide level to identify achievement trends and give guidance to Department staff providing technical assistance to districts. Reports of results are sent to districts and schools as well, to make use of as they wish. What are some ways that a district could make use of these assessment results?

In many cases, what is good for the state is good for a district as well. Because the objectives tested in the last three assessments (1981, 1983 and 1985) have remained almost totally consistent, it is possible to look at changes in districtwide results and draw valid conclusions about achievement trends.

These achievement trends are particularly meaningful if a district has matched the objectives covered in the assessment tests with their own curriculum. Because of the careful process used to determine the assessment objectives, the match between a district's curriculum and the objectives tested by the assessment will probably be quite good. It should not be taken for granted, however, as variations do occur among districts in the grade levels at which certain skills are expected to be mastered.

If a district's curriculum for grades K-4 includes most of the objectives tested in the fourth grade assessment's math computation subtest, for example, then one would expect to see improvements between the 1981 and 1985 fourth grade math computation scores. On the other hand, if most of the objectives tested in the eighth grade study skills subtest do not get covered in the district's curriculum until grade 9 or later, then no improvement in scores between 1981 and 1985 would be expected.

The question of expected achievement can be addressed in another way, too. The statewide Interpretive Panel process can be replicated in a district as well. This is the procedure used to determine whether scores are higher or lower than expected by the educators responsible for the students' in-school education. As a result of these determinations, it is possible to say that students' performance represents an area of strength or weakness. This process can be undertaken whether or not a formal matching activity comparing the assessment's objectives and the district's

continued, over

curriculum has been undertaken. More detailed information on how to conduct an interpretive panel process appears in an article below.

To Summarize . . .

Statewide assessment data should be considered supplementary to other districtwide test data available from standardized or locally developed tests. What is done with those test data can most likely be done with statewide assessment data as well: track achievement trends, identify student/curricular strengths and weaknesses, and so on.

If there is a good match between what is covered in the assessment test and what is included in the district's curriculum, there is even an argument to be made for letting the statewide assessment take the place of other reading and math tests administered to fourth and eighth graders. Because the Department of Education supports the Statewide Assessment Program in order to get the information it needs, it makes good sense for districts to make use of this "free" relevant test data.

Comparing Local and Statewide Expectations: The Interpretive Panel Process

Results from the Statewide Assessment provide general information about the basic skills achievement of Alaska's fourth and eighth graders. But the test results are not sufficient in and of themselves to allow meaningful interpretations to be made about that achievement. Since the 1981 assessment, statewide Interpretive Panels have been convened to provide the additional information needed to make statements of student strengths and weaknesses as shown by the test results. Local districts can use this process, as well, to add useful data to their local results.

PRACTICAL TIP

Interpretive panels can produce some very useful information, but they also produce a difficult public relations issue. Why? Because if the panelists set their expectations too low, the public will think the staff's goals are too low. But if the staff holds high expectations, there will be lots of weaknesses, making the quality of instruction look bad. How can this "Catch-22" be avoided? By viewing the Interpretive Panel as an information source more for educators than for the general public. Don't release the raw interpretive panel findings to the media. Instead, present the curricular and instructional actions that you'll be taking based on the panel information.

The process can be conducted using either questionnaires or an Apple II computer program available from the Department of Education. The computer program is probably best suited for small city and borough districts where all teachers at a grade level are in just one building. In larger districts or REAAs, where teachers are physically separated, use of the questionnaires is preferable since all input can be gathered by mail.

Regardless of the approach, Interpretive Panel members review each test item and make two determinations. First, they decide how low a percentage of students in the district could answer the item correctly and still leave the district satisfied with their performance. This is called the minimum level of performance. Next, they estimate the highest percentage of students that could realistically be expected to answer the item correctly. This is called the desired level of performance. Judgments are made without seeing students' test results. Districts might consider asking selected teachers to conduct the task at the same time as their students are taking the assessment test or immediately afterward.

Once final minimum and desired levels are computed for each test item, a comparison is made with the district's item p values (the percent of students answering each item correctly). When actual performance falls below the minimum, it is said to indicate a weakness. When performance exceeds the desired level, it is said to indicate a strength. Performance that falls between the minimum and desired levels is regarded as satisfactory.

Since it is important for Interpretive Panel members to know both the district's curriculum and the typical performance of students at the grades tested in the assessment, it is likely that most (but not necessarily all) panel members will be fourth and eighth grade teachers. If there are enough teachers on the panels, it is a good idea to ask half of the fourth grade panel members to review the math portion of the fourth grade test and the other half to review the reading portion and, similarly, half of the eighth grade panel to review each test section. If the numbers are too small, though--say less than ten persons--the same panel member can review both the reading and math sections. Experience with the statewide Interpretive Panel shows that panel members spend less than 30 minutes reviewing each test section.

While information about strengths, weaknesses and satisfactory performance is interesting in isolation, it becomes especially useful if a district has matched the assessment items with its curriculum. Then it is possible to make quite specific interpretations about the district's performance. For example, perhaps all of the identified weaknesses (according to Interpretive Panel results) were on objectives not included in the district's curriculum. This might suggest that the students are performing well enough compared to the curriculum; but it might be worth reviewing the curriculum to make certain that the missing skills were intentionally not covered. Alternately, though, a district may find that the students have not mastered skills which the district thought had been taught. Then it is time to review the instructional methods and materials used to teach the curriculum.

Examples of materials used with the 1985 Statewide Interpretive Panel, which districts might modify for their own use, appear on pages 10-A and 10-B.

Validating Test Scores

Many laypersons believe almost any reported statistic; if a number is printed in a newspaper or report, they think it must be "the truth." But educators know that some numbers--including test scores--may not accurately reflect the way things really are.

We have all read guidelines that remind us not to make decisions based on just one piece of information. Test scores, for example, should never be the sole criterion for an action. Instead, they might be corroborated by teacher judgments. Or results from one year might be corroborated by scores from a previous year. Or one test's results might be supported by results from a similar test given as a check on the first.

How might a district corroborate their statewide assessment scores? The form below shows a framework for an analysis that can be conducted to determine whether assessment results support or conflict with other evidence available to a district. It requires a comparison of assessment items with the district's curriculum objectives, an activity we have recommended elsewhere in this chapter. If the assessment is valid for a district, one would expect that there would be a higher percentage of items in the area of the form marked A than in the area marked B and, probably, than in the area marked C. Area D's percentage should be the lowest percentage of the four.

Items Measuring Objectives

	In Curriculum	Not in Curriculum
Assessment Items--Percent Correct	**A**	**B**
Assessment Items--Percent Incorrect	**C**	**D**

If a district has access to a computer and a staff member with expertise in statistical analysis, it might be worthwhile to compute a correlation between students' math assessment scores and their standardized math test scores, and between their two reading scores. Depending on the standardized test series, some of the assessment's reading subtests may be more appropriately correlated with language arts subtests.

It would be useful to determine the degree of overlap in objectives tested by the assessment and standardized test before interpreting the correlation results. If there is little overlap in objectives, we would not be concerned by low correlations between test scores. If many of the same objectives are included in both tests, though, we would expect higher correlations.

Just how high is a "higher correlation"? Unfortunately, there is no easy rule of thumb that can be used to answer that question because the answer depends on the number of students tested. If the same 100 students were tested on both the assessment test and the standardized test, a correlation of .19 between the two scores means that they are related to a statistically significant degree. But if only 20 students are tested using the two instruments, a correlation of .40 is needed to make the same statement. (Remember that correlations range from -1.0 to +1.0.)

So let's assume that we obtain a correlation of .30 between the district's fourth grade math assessment score and the standardized test math score; let's further assume that most of the same objectives are tested by both the assessment and the standardized test. If we have those two scores for 150 students, then a correlation of .30 means that the assessment results and standardized test results corroborate each other. If we have scores for just 15 students, though, we couldn't say that.

Unfortunately, in probably half of Alaska's districts, there are not sufficient numbers of students to warrant statistical analysis such as that described above. The formula used to compute a correlation coefficient includes a term representing the number of students tested. Because of the small number of students tested at any one grade in many Alaska districts, it is very difficult to attain high correlation coefficients. It would be a mistake to interpret a low correlation as evidence of test invalidity in such a case. It also means that little importance can be attached to correlation coefficients calculated on these small samples unless the coefficients are fairly substantial in size.

Score Comparisons: Appropriate and Inappropriate

Currently the State of Alaska does not report statewide assessment results for individual districts. Only results for the state as a whole are released. But across the nation there is a growing trend for district-by-district, and even school-by-school, information to be made public. Given this general trend, it is critical that at least one person in every district be thoroughly briefed on appropriate and inappropriate uses of publicly-reported district assessment scores so that this information can be disseminated to schools, community groups and media representatives within the district. What are those things we should all be aware of when looking at listings of individual districts' assessment scores? Should that ever happen with Alaska's Statewide Assessment scores:

- Remember that in small districts--say those with fewer than 30 students in a grade--district averages are unstable (that is, they vary from year to year) and therefore not very useful for drawing conclusions about the district.
- Also remember than in most districts in Alaska, the number of students from just two grades (fourth and eighth) isn't sufficient to be truly representative of the district as a whole.

continued, over

Score Comparisons, continued

- When looking at districts ranked according to average score, remember that differences in average score from one rank to another can be very small. Similarly, small differences in the percent correct scores of districts may represent very small raw score differences.
- Recognize that the objectives tested in the assessment may match the curriculum better in some districts than in others. This is happen stance when it occurs, since no individual districts' curricula were used as the basis for the tests. Instead, subject matter experts selected the objectives they felt were most important to test from a comprehensive listing in the Alaska Objectives and Items Bank (AOIB). But it is true, for example, that some districts may choose to present long division in April of the fourth grade year, or emphasize dictionary skills in ninth grade. In such cases, students in those districts would be at a disadvantage compared to districts where division was introduced in October of the fourth grade and dictionary skills were a seventh grade focus.
- It is inappropriate to compute an average of the fourth and eighth grade scores and use that single figure as a measure of the district's performance on the assessment. That would be an appropriate procedure only if the same students took both the fourth and eighth grade test--something which is obviously impossible. To put the two groups together and determine an average score would be like measuring the height of all boys and girls in a school and reporting that the average student's height was 4'5". Mathematically, that's a correct statement; but it is really meaningless when you stop to think about it. Similarly, there's no "average 4th/8th" grader, so an average assessment score is a mythical concept.
- Finally, recognize that rankings of districts on assessment scores (or any educational outcome measure, for that matter) presume that all students in Alaska have equal educational opportunities. While that is the premise on which all public education in the United States is based, it is much harder to achieve in practice than in theory--even in areas which are quite homogeneous. In Alaska, where the geographic and cultural diversity are extreme, it is even more difficult. This must not be used as an excuse to stop striving for equal opportunities; but to assume that they currently exist would be naive at best.

References

Anderson, Beverly L. (1985, Summer). State Testing and the Educational Community. Educational Measurement: Issues and Practice, 5, 22-26.

Considers the strength of the state testing movement nationally and proposes a scheme for looking at test characteristics and purposes. Useful background information for state/local integration.

Citron, Christine Hyde (1982, Winter). Competency Testing: Emergi.g Principles. Educational Measurement: Issues and Practice, 1, 10-11.

A discussion by a lawyer of four legal precedents that relate to statewide competency testing: (1) appropriate use of competency tests is constitutional; (2) there must be adquate notice; (3) competency tests may not carry forward the effects of past racial discrimination; and (4) a graduation test must reflect material taught.

Madaus, George F. (1982, Winter). Competency Testing: State and Local Level Responsibilities. Educational Measurement: Issues and Practice, 1, 12.

A short excerpt from a textbook chapter. Mentions nine specific SEA and LEA responsibilities/actions that are needed if a statewide testing program is to work satisfactorily.

Womer, Frank B. (1981). State-Level Testing: Where We Have Been May Not Tell Us Where We Are Going in New Directions for Testing and Measurement. Testing in the States: Beyond Accountability, Jossey-Bass, pp 1-12.

A brief look at state level testing from the 1920s to the present day followed by a discussion of the rationales behind current statelevel programs. Also discusses ten ways in which state programs are alike or differ.

Information for Interpretive Panel Members

Thank you for your willingness to participate in the Interpretive Panel process for the 1985 Alaska Statewide Assessment. It is this process that provides the only statewide indication of whether Alaska's students are meeting the expectations held for them by their teachers and administrators. We guarantee that the 30 minutes you spend in this activity will be extremely valuable.

There are four enclosures with these instructions; please take a moment to ensure that you received all four:

1. The Interpretation Form,
2. A postage-paid envelope for returning the form to Interwest,
3. The test for the appropriate grade level, and
4. A Personal Services Agreement necessary for payment of your honorarium.

If any of these four items is missing, please call Interwest immediately (503/223-3396, collect). Assuming all is in order, please read the directions below before proceeding. Refer to the numbered sections of the Interpretation Form as you read through the instructions below.

Detailed Instructions for the Interpretation Form

1. NAME--Print your name on this line.

2. PANEL--Note that the grade level and subject area you are to interpret is at the bottom of the sheet. (If the form does not match the test booklet you've been sent--blue for grade 4, green for grade 8--we have made a mistake; please call us immediately.) Note also that you will review only the math section or the reading section, not both.

3. ITEM CONTENT--This information is provided simply to help you keep your place in the review.

4. MINIMUM--In the box, write the percentage that represents the minimum **percent of students you feel should have answered this item correctly.** You are setting the lower limit to the expectations you have regarding the skill measured by the item. You would consider it unacceptable if student performance fell below this percentage. For example, you might feel that at least 75 percent of all students at this grade should have been able to answer the first item correctly. If so, you would place a 75 in the "minimum" column by the first item. If the students' actual performance is below the panel's average minimum, the item indicates a weakness.

5. DESIRED--In the box, write the percentage that represents the **percent of students you would like to see answer the item correctly.** You are setting the upper limit to the expectations you have regarding the skill measured by the item. You would be very pleased if the student performance exceeded this percentage. For example, you might feel that, while some students can't realistically be expected to answer the first item correctly, 90 percent could and thus 90 percent is a reasonable goal for which to strive. If so, you would place 90 in the "desired" column for the first item. Obviously, you would like to see everyone answer the item correctly but, realistically, that is a suitable goal for only a few very easy items. For some very difficult items, the desired percentage might be only 30 or 40 percent. If the students' actual performance is higher than the panel's average desired level, the item indicates a strength. (The gap between the minimum and the desired level forms the "satisfactory" range, the range where performance is acceptable, but has not reached the goal.)

Report your "minimum" and "desired" scores as a percent of students who should answer the item correctly; the number will be between 0 and 100. Most panelists in the past chose to round the scores to the nearest 5 percent (e.g., 75, 80, 85); feel free, however, to use an exact percentage whenever you think it more meaningful.

Base your decisions of minimum and desired performance on the expectations you hold for the students in your particular setting. Do not attempt to set levels for the entire state. Interpretive Panel members were selected from all across the state. If each panelist reflects the expectations of his or her own setting, the panel averages will be reasonably representative for the state as a whole. (Incidentally, don't forget that the students took the test in late March; your expectations should be those you hold for fourth or eighth graders in late March.)

---

6. "UNIMPORTANT" ITEMS--While the content of the tests was carefully selected, you might encounter an item which appears to you to measure a skill which is of absolutely no importance to a student's education. If so, refer to the box marked "Unimportant Items" and circle the number of the item in the box. For example, if you feel item 27 does not measure anything worthwhile, the number "27" in the box should be circled. Do not cite items that might be of lesser importance than most, nor items that are more appropriate at other grades. Mark only those items which relate to a skill which is of virtually no value.

Should you have questions about the meaning of "minimum" and "desired," how to complete the form, or anything else about this task, use the Assessment Hotline (503/223-3396, collect) to obtain further information. When you have completed the form, use the enclosed postage-paid envelope to return it to us. Discard the test. Send your form back as soon as possible; your envelope must be postmarked no later than April 22nd for you to receive your honorarium. (Should anything--such as slow mail--make this deadline impossible, please call us collect to let us know of the delay.)

The Interpretive Panel process is perhaps the most important step in making the assessment information useful to Alaska's educational decision makers. Your participation is greatly appreciated.

10-A

ALASKA ASSESSMENT PROGRAM

INTERPRETIVE PANEL FORM 1

	minimum	desired
1. Dog team races		
2. Dollars-question		
3. Dollars-method		
4. Tackle shop-equation		
5. Tackle shop-fact		
6. Balloons		
7. Doghouse-question		
8. Doghouse-equation		
9. Sitka-method		
10. Sitka- fact		
11. Add, no regroup		
12. Add, regroup		
13. Add decimals		
14. Subtract, no regroup		
15. Subtract, regroup		
16. Subtract, regroup		
17. Subtract decimals		
18. Multiply, no regroup		
19. Multiply, regroup		
20. Multiply, regroup		
21. Division, no remainder		
22. Division w/remainder		
23. Words for number		
24. Sequence		
25. Place value		
26. Rounding-tens		
27. Rounding-hundreds		
28. Even/odd		
29. [illegible]ords for time		
[illegible] from clock		
31. Metric unit of length		
32. Customary unit of weight		
33. Meters in a kilometer		
34. Feet in a yard		
35. Multiplying by zero		
36. Identify fraction		
37. Words for fraction		

"Unimportant" Items

1 2 3 4 5 6 7 9 9 10
11 12 13 14 15 16 [illegible] 19 19 20
21 22 23 24 25 26 2[illegible] 28 29 30
31 32 33 34 35 36 37

4th grade Math

Name

MATHEMATICS APPLICATION

Application items test problem solving skills: identifying facts, translating written information into math equations, choosing the correct approach to solve a problem, determining which data are needed to solve a problem, and so on. Items generally consist of a short story problem, with one or two relevant questions.

GRADE 4

Summary of Performance
3 STRENGTHS 3 WEAKNESSES

Item Performance

Item 2 Understanding the Problem

Given a word problem requiring addition, 87.2% could answer the question "What does this problem ask you to find?" Range was 66.3 to 84.4. *A strength.*

Remember that the "range" indicated in each case refers to the span between the MINIMUM level and DESIRED level. In Item 2, for example, the MINIMUM level was 66.3%, while DESIRED was 84%. The 87.2% performance level was above this range, indicating a strength.

Item 3 How to Solve the Problem

In Part 2 of the same problem, 91.8% knew how to use addition to find the total number of marbles owned by three children. Range was 69.7 to 85.1. *A strength.*

Item 7 Understanding the Problem

In Part 1 of another two-part problem, 73.6% recognized that they were asked to find the area of a shed, rather than the height, width, or perimeter. Range was 40.9 to 71.4. *A strength.*

Item [illegible] Solve the Problem

[illegible] second part of the shed problem, 47.6% knew which formula (4 x 3) would give them the area. Range was 48.9 to 74.3. *A weakness.*

Item 9 How to Solve the Problem

Items 9 and 10 were based on a short word problem about Anchorage. Fewer than half the students—44.4%—knew how to use subtraction to determine the year in which Anchorage had become a city. Range was 48.2 to 74.3. *A weakness.*

Item 10 Identifying Facts

Item 10 proved difficult as well. Only 36.7% could give the population of Anchorage when it became a city, although this figure was directly stated in the problem. Range was 50.1 to 74.4. *A weakness.*

Comments

Notice that students did best on the two-part problem involving addition (Items 2 and 3). In the problems on area (Items 7 and 8), students knew what was asked for—many did not know it was necessary to multiply length times width to find area. Problems with Items 9 and 10 may have been caused by simply reading—since a number of students who responded correctly to these items seem to have confused the figures for each population.

GRADE 8

Summary of Performance
1 STRENGTH 5 WEAKNESSES

Item Performance

Item 2 Averages

This item required students to find the average of three figures involving decimals. As answered correctly, range was 65.7 to 90.4. *A weakness.*

Item 3 How to Solve the Problem

In the single best math performance by either grade on any item, 97.4% correctly identified adding as the way to solve a short word problem on the number of people visiting the Mt. Hood Glacier. Range was 88.5 to 96.9—relatively high, but performance was still better. *A strength.*

Item 12 How to Solve the Problem

53.6% knew how to solve this two-step word problem using subtraction. Range was 37 to 84. *A weakness.*

Item 13 Missing Data

63.2% knew which missing data were needed to solve a word problem on averages. Satisfactory range was just a shade higher: 63.6 to 87. *A weakness.*

Item 15 Unneeded Data

In this turnabout version of Item 14, only 49.4% could tell which data were NOT needed to solve a word problem requiring multiplication. Range was 58.6 to 82.7. *A weakness.*

Item 16 How to Solve the Problem

In Part 2 of the Item 15 word problem, 54.6% could select the correct equation to solve the problem. Range was 57 to 83.6. *A weakness.*

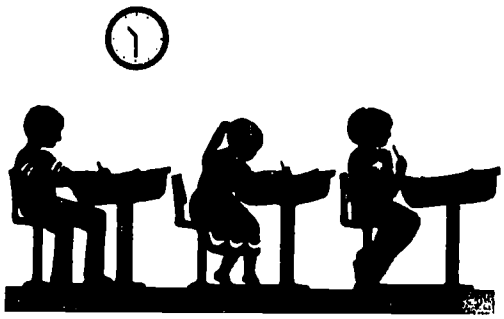
Comments

For the most part, eighth graders' performance on math application items fell within the satisfactory range, however, there were problems with calculating averages and solving word problems. In particular, eighth graders had trouble determining what additional data were needed to solve a problem, or which data were extra (Items 14, 15). Areas where students performed satisfactorily include:

- Identifying facts in a short word problem;
- Determining the right equation to solve a problem;
- Applying labels (meters, liters) correctly; and
- Calculating perimeter.

ASSESSMENT HANDBOOK

A Practical Guide for Assessing Alaska's Students



Glossary and Index

- **Definitions of important assessment terms**
- **Where in the Assessment Handbook a discussion of each concept can be found**

A Glossary of Important Terms

The list below contains definitions for a number of important assessment terms. While the terms may be used throughout the Assessment Handbook, the chapter number following the definition is where the topic is emphasized. If no chapter number appears, the topic is not a major topic in the Handbook, although an understanding of the topic may be helpful as you study other assessment issues.

A through E

- **achievement test:** A test that measures the extent to which a student has acquired certain information or mastered certain skills. Chapter 5
- **alignment:** The process of assuring that curriculum, instruction and testing all match each other, and that communication among educators and administrators at all levels within a district is open and functional. Chapter 1
- **aptitude test:** A test that measures a person's ability to learn or develop proficiency in some particular area if appropriate education or training is provided. Chapter 5
- **cohort:** A group of individuals who are comparable on some dimension: for example, students at the same grade level within a district or state. Chapter 6
- **cohort analysis:** Following the same group of students across grades, for example, to track their achievement from one year to another. Chapter 6
- **co-normed:** Two or more tests that are normed with the same group of students. Chapter 5
- **content validity:** The extent to which a test matches the curriculum objectives and subject content of a given program. Chapter 5

- **correlation coefficient:** A measure of the degree to reliability between two sets of measures for the same group of individuals. Correlation coefficients range from 0.00, indicating a complete absence of relationship, to +1.00 and -1.00, indicating a perfect positive or negative correspondence. Chapter 10

- **criterion referenced test (CRT):** A test that is designed to provide information on the specific knowledge or skills possessed by a student. The scores on a criterion referenced test have meaning in relation to what the student knows or can do. Chapter 2

- **educational significance:** A judgment that test performance, or the difference in test performance by separate groups, is meaningful or important in practical terms. This term is often contrasted with statistical significance (see below).

- **empirical norm dates:** The actual dates on which a test publisher tested the students in the norm group. Publishers recommend these dates to schools as the dates that should be used for administering the test. Testing at times other than the empirical norm dates means that students may have received more or less instruction than the norm group. Chapter 2

F through N

- **formative evaluation:** An evaluation conducted at a time when a program can still be modified. The primary purpose of such an evaluation is to collect information that will help improve the program. Chapter 2

- **grade equivalent score (GE):** The grade level for which a given score is the real or estimated average. Chapter 5

- **item analysis:** The process of evaluating individual test items to assure their quality with respect to certain characteristics. Item analysis involves determining such factors as the difficulty value and discriminating power of the item. All such characteristics are then used to judge the overall quality of the item.

- **normal curve equivalent (NCE):** A measurement scale developed for Title I (Chapter 1) evaluation requirements. The scale ranges from 1 to 99, with units equal in size across the score range. The equivalence of units makes it possible to average scores across groups and to aggregate results across tests.

- **norm group:** The sample of students to whom a test has been given in order to estimate how well the student population in general would perform on the measure. A norm group should be as representative as possible of the variation expected within the general population. Key dimensions to be represented in a norm group include ethnicity, socioeconomic status, size of school system, location of system (urban, rural or suburban), public vs. non-public schools, and geographical region of the country. Chapter 3

- **norm referenced test (NRT):** A test that is designed to provide information on how well a student performs in comparison to other students. The scores on a norm referenced test have meaning in terms of their relation to the scores made by an external reference group. Chapter 2

PRACTICAL TIP

It's not hard to find definitions and explanations of measurement terms and concepts, but it is hard to find ones that are understandable. You'll find that statistics and measurement textbooks are frequently obtuse, and usually assume you have a deep passion for formulas and Greek letters. Use them as a last resort, but first direct your attention to the "practical" references given at the end of each chapter of the Assessment Handbook.

- **norms tables:** Tables presented in test manuals or available from test publishers that show the relationship of different types of scores to one another (e.g., raw scores to percentiles). Tables are usually provided for each test level and time of testing (norms dates) as well as by grade level of the student tested. Chapter 5

O through Z

- **out-of-level testing:** Administering a test at a level below or above the one generally recommended for a student based on his or her grade level. Such testing is done to accommodate the ability levels of students who are either much above or much below the average of students their age and thus would not be able to demonstrate the knowledge and skills they possess. Chapter 3

- **percentile rank:** An indication of a student's standing in comparison with all students in the norm group who took the same test. Percentile

ranks range from a low of 1 to a high of 99. A percentile rank stands for the percentage of students who obtained scores equal to or less than a given score. Chapter 5

- **p value:** An index which signifies the percentage of examinees who answered a test item correctly. Chapter 10

- **raw score:** The number of test items answered correctly by a student. Because different tests have different numbers of items, raw scores cannot be compared from one test to another. Chapter 6

- **reliability:** The extent to which a test can be depended upon to provide consistent, unambiguous information. Reliability is usually reported as a correlation coefficient, with the closer the coefficient is to +1.00, the higher the reliability. Types of reliability commonly reported for tests include test-retest, alternate forms, split half and Kuder-Richardson (KR) 20. Chapter 5

- **scale score:** A score that expresses the results of a particular test for all forms and levels on a single common scale. Scale scores allow comparisons to be made from grade to grade or level to level of a test. Chapter 5

- **standard score:** A general term referring to any of several types of "transformed" scores. Raw scores are expressed in terms of standard scores for reasons of convenience, comparability and ease of interpretation. For example, the raw scores of two tests can be expressed in comparable terms by using standard scores. Chapter 6

- **standardized test:** A commercially published test designed to provide a systematic sample of individual performance. The test is administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Chapter 3

- **stanine:** A standard score scale that ranges from a low of 1 to a high of 9, with a specified percentage of cases falling into each category. Chapter 6

- **statistical significance:** A judgment, based on the application of statistical calculation, that a certain test score or the difference in scores between separate groups are "really" different--that is, not just apparently different because of chance fluctuations. While statistical significance gives the appearance of scientific truth, it must be understood that results of statistical analyses are very dependent on the number of students tested. The smaller the number of scores analyzed, the bigger the difference is required for it to be statistically significant. For this reason, many persons talk about both statistical and educational significance when referring to test scores.

- **summative evaluation:** An evaluation conducted at a time when a summary judgment is to be made about a program. The primary purpose of such an evaluation is to collect information that can be used to determine whether a program should be retained or deleted. Chapter 2

- **survey battery:** A group of several standardized tests administered together. Such tests typically cover a broad range of basic skills content. Chapter 5

- **testwiseness:** The possession of skills independent of subject matter knowledge that make it possible for students to achieve better test scores. Such skills can be taught and will result in small but consistent improvement in test scores. Chapter 3

- **validity:** The characteristic of a test that refers to whether the items in the instrument are a fair measure of the content or construct the test says it is measuring. There are various types of validity. Content validity is of major importance in achievement tests; predictive validity is a critical characteristic of aptitude or ability tests; and construct validity is a requirement for many psychological tests.

A FINAL NOTE

There is much to know about educational testing that can't be covered in detail in the limited space available in this handbook. But that doesn't mean you don't have any help available. The Division of Educational Program Support of the Alaska Department of Education has trained staff and contractors who can provide technical assistance in all areas of student assessment. You can write them at Pouch F, Juneau, Alaska 99811 or call (907)465-2900.