DOCUMENT RESUME

ED 300 414 TM 012 410

TITLE Assessment Handbook: A Guide for Assessing Illinois' Students.
INSTITUTION Illinois State Board of Education, Springfield.
PUB DATE 88
NOTE 106p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Achievement Tests; Curriculum Development; Educational Assessment; Elementary Secondary Education; Evaluation Methods; Evaluation Utilization; Recordkeeping; Resource Materials; *School Districts; *Standardized Tests; *State Programs; *Student Evaluation; Test Items
IDENTIFIERS *Illinois

ABSTRACT

This manual, which is provided as a resource for Illinois School Districts and which incorporates some of the materials in the Alaska Department of Education's Assessment Handbook, is designed to meet the requirements of Public Act 84-126 and to help practitioners develop student assessment plans of local usefulness. Areas covered in the guidelines include the nature of assessment and a comprehensive assessment system, involvement of staff and other constituent groups, types of assessment, alignment of assessment with curriculum and instruction, selection of standardized achievement tests, construction of local assessment procedures, administration of assessment programs, keeping records, and reporting assessment results. Assessment planning roles for teachers, administrators, curriculum coordinators, school board members, and parents are outlined. Other important considerations addressed include maximizing validity and reliability, avoiding bias, using multiple approaches, assessing different cognitive levels, using assessment procedures for multiple purposes, and developing an assessment schedule. Appendices include a glossary of assessment terms, a list of contact persons for proposals for funding for development and dissemination of effective assessment practices, and guidelines for improving test questions. (TJH)

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students

AN ASSESSMENT HANDBOOK FOR ILLINOIS SCHOOLS

Illinois State Board of Education

1988

(Printed by Authority of the State of Illinois)
(65483—5M—4-88)

Foreword

This manual is provided as a resource for Illinois school districts. The manual is three-hole punched so that it can be kept in a loose-leaf notebook and used as needed for local planning and development. Individual sections can be removed and duplicated. Additional materials can be inserted. In the future, we will issue additional chapters as needed.

We appreciate the graciousness of the Alaska Department of Education in allowing us to adapt and use materials from their *Assessment Handbook.* Those materials, which are acknowledged in individual chapters, have been of great assistance in the preparation of this manual.

If you have questions about the information in this manual, contact the Student Assessment Section at 217/782-4823.

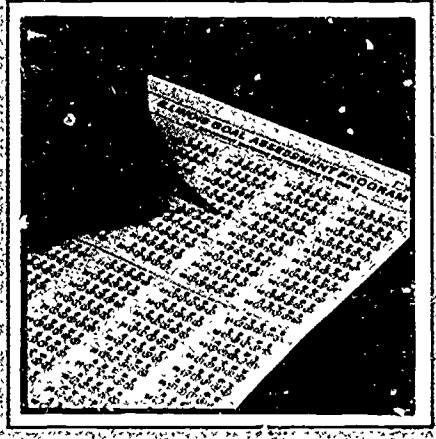
Ted Sanders
State Superintendent of Education

TABLE OF CONTENTS

General Appendix

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 1

This handbook is being distributed to help Illinois school districts meet the requirements of P.A. 84-126, while at the same time developing student assessment plans of high quality that produce information which is useful locally. Assessment is a critical element of the objectives-assessment-school improvement cycle. This handbook includes information for districts to use in developing effective assessment systems.

What Is Assessment?*

The terms *test, assessment,* and *evaluation* are frequently used interchangeably, but, in fact, have important differences.

Test, the narrowest of the terms, usually refers to a specific set of questions that will be administered to an individual or to all members of a group It is tangible and structured and can be administered within a relatively limited period of time.

Assessment is more encompassing. Testing is part of assessment, but it is only one measurement approach. Assessment may also include other procedures such as rating scales, observation of student performance, individual interviews, or reviews of a student's background or previous performance. Assessment may refer to groups or individuals. Group assessment may involve administering subsets of items to different samples of students and reporting the results for groups but not individuals. In addition, assessment often refers to a planned program of assessment. This handbook is about both assessment and testing.

Evaluation, as the word itself suggests, refers to making a value judgment about the implications of assessment data. This process is necessary for school improvement planning. While assessment involves obtaining performance data through a variety of means, evaluation goes a step further – interpreting the data from an informed perspective. That perspective should be informed by other information as well – for example, information about instructional content, community context, school climate, and dropout rate This handbook includes some material on the interpretation of assessment results, but, for the most part, evaluation is not included here

*Section adapted from Alaska *Assessment Handbook.*

In summary, testing provides one isolated glimpse—analogous to taking a picture with a camera—of how a student or group of students is performing on specific skills at a specific time. Assessment provides more comprehensive data on student performance through several administrations of test batteries or through various other data-gathering approaches. Evaluation produces value judgments about the results provided through assessment.

A word of caution. 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 Is a Comprehensive Assessment System?

A comprehensive assessment system is a coordinated plan for periodically monitoring the progress of a district's or school's students at multiple grade levels in a variety of subject areas. It specifies the procedures that will be used for assessment; indicates when and how those procedures will be administered; and describes plans for processing, interpreting, and using the resulting information. A good comprehensive assessment system includes:

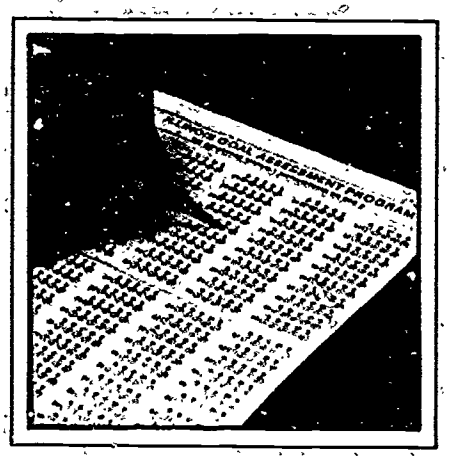
- A schedule for assessing students throughout the school year and at all grade levels.
- Multiple types of assessment procedures (e.g., norm-referenced tests, criterion-referenced tests, locally developed performance rating scales) that are used appropriately.
- Standards that protect the quality of student assessment. At a minimum, those standards refer to the reliability of the test/other assessment procedure administration and the validity of the interpretation and use of resulting data (AERA, APA, and NCME, 1985).
- Provisions for collecting other relevant information (for example, analysis of the instructional process, judgments about local conditions/needs, and contextual/system variables such as mobility and dropout rates) that can be used to supplement achievement data during decision making.
- Plans for processing and using the assessment results. Such plans can help districts design assessment systems that meet information needs, are efficient, and minimize the paperwork burden on staff.

How Does a District Develop an Assessment System?

The process that a district uses to plan its assessment program is critical to the program's success. Naturally, that process helps determine the quality of the system and its appropriateness for local purposes. Another very important potential influence of the development process is that it elicits crucial support from people who will approve the system, allocate resources to it, and implement it. Without that support, the system may not have a chance. One strategy for generating that support—involving constituent groups in the program's development—is the topic of the next chapter.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 2

Different Groups—Different Roles

To play a significant role in education, an assessment program must have the support of both the taxpaying public which 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, students, and others. District administrators may develop that credibility 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 planning, implementing, maintaining, and evaluating assessment programs.

To the extent possible, those who are affected by assessment results (district and school administrators, teachers, students, parents, and other community representatives) should participate in the assessment's design. Each group offers a unique and vital perspective on which skills are most important to assess, how to assess them, and especially how to use the results.

Several issues for discussion among these assessment decision makers include:

1. What should be the major purposes/functions of assessment?
2. What types of assessment procedures are most appropriate for various goals/objectives?
3. When should assessment occur (both time of year and frequency)?
4. How should results be reported? What data do various groups want and need?
5. How should assessment results be used?
6. What do results indicate about student performance?

*Chapter adapted from Alaska *Assessment Handbook.*

7. What revisions are needed in local objectives, student expectations, teaching-learning activities, or assessment procedures?

8. How can the assessment system be improved (for example, by modifying the assessment approaches, processing procedures, or reporting)?

9. How can the assessment program better serve its users?

Not all groups need to participate in discussions about all questions. Different groups contribute different strengths, and these strengths should influence how various constituent groups are involved. The chart in the Chapter Appendix shows how the different responsibilities might be distributed across groups.

Committees Serve a Variety of Needs

Depending on a district's particular needs, several different types of advisory groups should participate in long-range assessment planning. These could include an overall committee, task forces, and an interpretive panel. The composition and function of each group is discussed in the following paragraphs.

Overall Committee

This committee will have general responsibility for developing the assessment system. It should be broadly representative of the district. The committee should probably include one or two district administrators who will be involved in system implementation or the use of resulting data (e.g., testing or curriculum directors), one or two building-level administrators, a number of teachers from different grade levels and subject areas, and perhaps teachers association or union representatives and community (e.g., school board) members.

This committee can form an important link between local teachers and administrators and, depending on local needs, the link may be broadened to include others. Throughout the planning process, the committee should inform others of its work and invite their suggestions. The committee will probably want to involve other staff more directly by establishing special-purpose task forces.

One of the committee's first tasks will be to develop a framework for the assessment system. That framework should include the functions or purposes of the system and the types of assessment procedures to be used. (Several functions or purposes the committees may want to consider are listed in the Chapter Appendix. Types of assessment procedures are discussed in Chapter 3.) The framework can then be used—by the committee as a whole or by task forces or subcommittees—to guide the development of more specific assessment plans.

Task Forces

The number, composition and major functions of the task forces will vary according to local factors such as development needs, district size, and type of district (elementary, secondary, or unit). The functions of the task forces may include selecting or developing assessment procedures for various learning areas/grade levels/goals or designing specific assessment system elements such as procedures for processing or reporting the assessment data.

Interpretive Panel

An interpretive panel can improve the use of assessment results by reviewing them and suggesting what the results reveal about student performance. The interpretive panel can examine the numerous complex factors affecting student performance and help audiences understand the relationship among curriculum, instruction, and assessment. They can provide valuable guidance about school improvement.

The School Board's Role

A well-informed school board can be one of the best allies of any assessment program. Proactive administrators keep their boards closely involved throughout the planning and implementation of assessment programs. Regularly scheduled board review can help ensure that testing practices remain responsive to a district's changing needs.

At each stage of the planning, implementation and review of an assessment program, a skillful administrator provides the board with relevant decision-making information. The information is more likely to be used if it is timely, complete, easily understood, and targeted to the decisions at hand.

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

- the potential purposes of assessment,
- the limitations of assessment,
- state assessment requirements,
- components of an effective assessment program,
- estimated costs for various assessment options, and
- proposed procedures for test selection or development.

After an assessment program has been implemented, the board should be kept informed about it. 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 for questions/comments from the public.

The assessment program should be reviewed regularly— whenever students do not meet locally defined expectations in one or more learning areas (as indicated in School Improvement Plans) or according to a pre-established schedule. The review should include a staff analysis of the appropriateness and usefulness of the assessment procedures. Also, various audiences can be surveyed about their perceptions of the program's effectiveness. Staff recommendations for changes in the assessment program can be presented to the board, along with survey results. If the board members have been provided with appropriate information since the beginning of the assessment program, decisions made at this point should be especially sound.

Practical Tip

Why does assessment generate so much controversy? One important reason is that people (including teachers) feel they were not part of the process, that decisions were made without their knowledge, leaving them with only a "take it or leave it" option. Some fear that inappropriate assessment procedures will produce misleading results that make them look bad. When all parties are involved from the beginning, assessment programs should be better and the amount of criticism reduced.

Special Considerations

Members of all committees, task forces, and panels should be selected carefully. Whether they are asked to serve or selected from a pool of volunteers, they should be people who are:

- interested in the development (or interpretive) task,
- knowledgeable about the educational program,
- supported and respected by district teachers and administrators, and
- able to devote time to the task.

Developing a comprehensive assessment system is time-consuming. However, it is a very important task which must be done with care. District administrators may need to identify strategies that make it easier for school personnel to participate. For example, they might schedule meetings during in-service or institute days, hire substitutes, ask other teachers to cover participants' classes (and perhaps reward them for doing so), or pay staff for working during the summer.

References

Belleville Public Schools, District #118. (1986). *Instructional Monitoring System. A Manual for School Districts Interested in Replicating This System.* (Available at Educational Service Centers.)

Includes discussion of teachers' involvement in system development and of continuing communication with them.

Woodstock Community Unit District #200. (1986). *Student Assessment System.* (Available at Educational Service Centers.)

Includes description of staff involvement in assessment system planning and implementation.

Zion Elementary School District #6. (1986). *An Aligned Instructional Goal Assessment Management System for Student Learner Outcomes.* (Available at Educational Service Centers.)

Describes rather extensively the process of working with staff to develop an assessment system.

Appendix A: Chapter 2
Suggested Roles of Constituent Groups in Assessment Planning

GROUP	Determination of Assessment Functions/ Purposes	Selection or Development of Assessment Procedures	Implementation of Assessment System	Interpretation of Assessment Results	Evaluation of Assessment System
TEACHERS	Propose assessment purposes and priorities	Review standardized tests/write items/ develop other assessment procedures	Administer assessment	Help interpret results; suggest appropriate school improvements	Review system and suggest modifications
ADMINISTRATORS	Review proposed assessment purposes and priorities; develop proposal for school board	Oversee assessment development	Oversee assessment administration and reporting of results	Review school and district results; develop local improvement plans and report results	Review and modify assessment system
CURRICULUM COORDINATOR/ OTHER INVOLVED STAFF	Review proposed assessment purposes	Review standardized tests/participate in other asssessment development	Assist with assessment administration; handle processing and reporting	Help interpret results and advise school improvement decisions	Review system and suggest modifications
SCHOOL BOARD MEMBERS	Critique proposed assessment priorities	Approve proposed general assessment procedures		Review assessment results and improvement plans	Review system and approve suggested modifications
PARENTS/OTHER COMMUNITY MEMBERS	Suggest assessment priorities			Review assessment results	Review system and suggest modifications

Appendix B: Chapter 2
Assessment Purposes/Functions

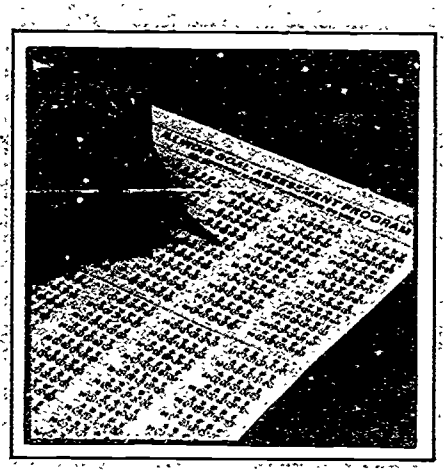
Early in the process of developing an assessment system, local planners need to decide what functions the system should serve, a decision which directs the remainder of the process and provides a basis for formal board policy statements regarding assessment. Some assessment functions/purposes that may be considered include:

- To meet the requirements of P.A. 84-126 (assessing third, sixth, eighth, and eleventh grade students on local learning objectives in the six fundamental learning areas annually).
- To help assure that students meet local objectives by monitoring their progress more frequently.
- To meet accountability expectations by collecting achievement data that permit comparisons of local students with national norms.
- To make better school improvement decisions.
- To monitor student progress on classroom-level instructional objectives.
- To diagnose student needs.
- To place students in special classes or instructional groups.
- To provide information for career or psychological guidance.
- To obtain college entrance examination data.
- To evaluate special programs.

Decision makers should consider these and other potential assessment functions carefully. They may want to limit the purposes of a districtwide assessment system, for example, to focus on the first four functions listed above and leave the others to the discretion of teachers, guidance counselors, or other local personnel. However, they may also want to consider other functions in order to minimize the number of assessment procedures that must be administered by selecting those which can serve multiple functions—for example, a norm-referenced test that can help assess local objectives and meet accountability expectations and Chapter 1 reporting requirements.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 3

- Publisher's Standardized Shelf Tests
- Publisher's Customized Tests
- Publisher's Textbook Tests
- District's Locally Developed Tests
- Uniform Procedures Requiring Written Performance by Students Scored According to a Uniform Rating Scale
- Uniform Procedures Requiring Other (Nonwritten) Performance by Students Scored According to a Uniform Rating Scale

After identifying the functions of assessment, committees should decide what general types of assessment procedures are most appropriate for each. Later, task forces can select or develop more specific assessment plans. This chapter describes various types of assessment procedures; the uses, strengths, and weaknesses of each, and the special considerations they require (summarized in Table 3.1). The assessment procedures discussed here correspond with those on the Learning Assessment Plan (LAP) form (ISBE 41-78):

Publisher's standardized shelf tests,
Publisher's customized tests,
Publisher's textbook tests,
District's locally developed tests,
Uniform procedures requiring written performance by students scored according to a uniform rating scale,
Uniform procedures requiring other (nonwritten) performance by students scored according to a uniform rating scale.

The first four procedures refer to what is commonly known as forced-choice testing approaches. That is, students take paper and pencil tests and select the correct response from two or more alternatives (e.g., multiple-choice, true-false, or matching test items) or supply a word or short phrase to answer a question or complete a statement. Students' scores are likely to be comparable regardless of which teachers administer and score the test. The last two types require student performance. Here, uniformity is emphasized in order to reduce the variance which normally occurs when teachers judge student performance, but do not use the same criteria or standards. Without uniformity, the scores of students with different teachers cannot be summarized together or compared. Due to the need for district-wide student assessment data, all teachers in a district must use the same instructions to elicit student peformance and the same criteria and standards for rating the performance.

Publisher's Standardized Shelf Tests

Most tests that districts know as standardized achievement tests are publisher's standardized shelf tests. (Examples are listed in the instructions for the LAP forms.) Published norm-referenced and criterion-referenced tests are both in this classification.

Publisher's standardized shelf tests can serve many different purposes. Sometimes selected items or subtests can be used to assess student achievement of local objectives. Generally, data from these tests will help meet accountability needs, inform school improvement, and meet reporting requirements for special programs. Sometimes these tests can also provide information for diagnosing student needs or making student placement decisions.

Local planners who use published tests to assess local objectives may decide to select individual items, eventually clustering the items to measure particular objectives or sets of objectives. Or, they may decide to adopt particular subtests. Occasionally, an entire published test may be appropriate for assessing local objectives.

A major advantage of publisher's standardized shelf tests is their multiple potential uses, but the tests have several other advantages also. They are likely to have been professionally constructed. They have probably been reviewed carefully, pilot-tested, and analyzed for bias. Also, information about reliability and validity are usually available in technical manuals and other sources. Item-difficulty data, which can be very useful in deciding how to use an item and in setting student expectations, may also be readily available. Scoring and reporting services can be obtained.

These tests must be used carefully, however. Local staff should screen the items closely to determine their appropriateness for measuring student achievement of local objectives. Test publishers may provide information to help with this task (for example, lists of items which assess particular objectives). However, teachers should still examine the items to find out whether they agree with the test publisher's interpretation of the objective and of what the items measure. And, local planning groups need to make sure that the test results can be aggregated as needed (that the publisher can provide student performance data for the specific clusters of items that measure particular local objectives) and that the district orders a scoring package which will include that information.

Publisher's Customized Tests

Customized tests are tests which publishers tailor to local needs. The publishers often maintain banks of objectives and test items to measure the objectives. A district selects objectives and the publisher identifies relevant items and assembles them into a test.

Customized tests can be particularly useful for assessing local learning objectives. They have other advantages also, many of which are similar to those of standardized shelf tests. Generally, the items have been professionally developed. Data may be available about validity, reliability, and item difficulty. Scoring and reporting services are available.

However, districts should be cautious about using this assessment option. The tests (and related scoring and reporting services) can be very expensive. The objectives that are selected from the publisher's bank may not be sufficiently aligned with local objectives. Local personnel should examine individual items and consider their appropriateness for measuring specific local objectives, as well as the level of knowledge assessed. Like other standardized published tests, customized tests are generally limited to forced-choice (e.g., multiple-choice or true-false) items that can be readily scored by machine. They may not be valid for assessing some skills in areas such as writing, speaking, fine arts, or physical development.

Publisher's Textbook Tests

Textbook tests are the tests that accompany textbooks. They may be particularly appropriate for assessing student progress on local objectives, including classroom-level instructional objectives. They have several advantages: They are readily available, usually, closely aligned with instruction; and inexpensive. Districts do not have to spend additional money to purchase them or time to develop them.

Before deciding to use textbook tests to measure local objectives, local planners should take several factors into consideration. The test items should be aligned with local objectives. (Since districts will report information about student achievement of objectives, aligning tests only with instruction—i.e., use of the textbook—may not be sufficient.) Districts may need to ask textbook publishers for additional information about test items. For example, what is known about the items' reliability, validity, and difficulty level? Have the items been pilot-tested or reviewed for bias? Districts may also need to establish procedures for duplicating the items and scoring them uniformly.

District's Locally Developed Tests

An attractive alternative, locally developed tests may be the most appropriate type of procedure for assessing many local objectives. Since locally developed tests will be administered on a smaller scale than publishers' tests, they may permit more opportunities for such alternatives as having students listen to audio recordings as they respond to multiple-choice questions about music, using videotapes with questions about dance or drama, or simulating scientific experiments on computers. Also, some local audiences may consider these tests more credible and use the results more.

Developing local tests is a very ambitious undertaking, however. The process is very time-consuming. Writing good assessment items is difficult. Districts should develop several items for each local objective. Local staff must address issues such as reliability, validity, nondiscrimination and item difficulty. They will also need to develop processes for printing, scoring, and reporting.

Uniform Procedures Requiring Written Performance by Students Scored According to a Uniform Rating Scale

Students' writing skills may be tested by an assessment procedure in which students are given a standardized writing assignment and trained raters assess the results using a common scale. (The Illinois Writing Assessment Program illustrates this approach. It has been used widely and tested extensively. Standardized prompts and uniform rating scales exist. See *Write on, Illinois!*, ISBE, 1987.) This type of assessment may represent the most effective procedure for assessing students' writing skills. Furthermore, it can be useful instructionally.

The procedure is somewhat demanding and time consuming. Local planners will need to obtain or develop writing prompts and to ensure that raters are trained. (Fortunately, resources and technical assistance are available in Illinois through the Educational Service Centers.) The rating process should be monitored periodically to determine that it is being applied uniformly.

Uniform Procedures Requiring Other (Nonwritten) Performance by Students Scored According to a Uniform Rating Scale

In performance assessment procedures, students demonstrate their achievement in a particular area by performing a specified task. For example, they might draw a sketch, play a musical instrument, or run a 100-yard dash. The specific assignment should be described in the local assessment procedures, as well as instructions for rating the performance uniformly. (As with all locally developed assessment procedures, districts should maintain centrally located files in which performance stimuli/assignments and uniform rating scales are described thoroughly.)

The performance might be live or recorded. For example, a student might perform a dance or play, conduct a laboratory experiment, or compete in an athletic contest as a teacher rated his or her performance. Or, the performance might be videotaped so that the teacher could view it later. Other forms of recorded performance might include portfolios of sketches, photographs, or choreography; audio recordings of speeches or musical performance; and works of art such as paintings or sculptures.

A major advantage of this type of procedure is its effectiveness in assessing some objectives, especially in the fine arts and physical development and health. Many objectives that refer to student performance cannot be assessed validly using paper-and-pencil tests.

Performance assessment procedures must, of course, be used uniformly. Standardized directions for eliciting student performance and scales for rating that performance must be developed. Teachers must be trained in their use. Without uniform administration and scoring practices, the results will not be useful to anyone other than individual teachers. The aggregation of data across students or schools—which, for example, is required for assessing local status on learning objectives—is not justified. The additional effort this requires, however, may be far outweighed by the meaningfulness and utility of the results.

References

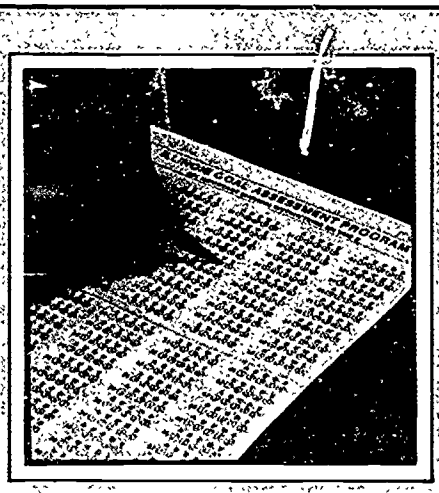
Illinois State Board of Education (1987). *Write on, Illinois!* Springfield, IL. Illinois State Board of Education. Two volumes of this document are available. The second is an updated and revised version of the first. (Two copies of the second volume have been distributed by Educational Service Centers to each Illinois school district.) Both are users' guides for scoring student essays, the fifth type of assessment procedure described in this chapter.

Table 3.1
Advantages of and Cautions Regarding Various Types of Assessment Procedures

Type of Procedure	Advantages	Cautions
Publisher's standardized shelf tests	Useful for many purposes. Professionally constructed, piloted, and reviewed for bias. Validity, reliability and item-difficulty data probably available. Scoring and reporting services available.	Require careful examination of alignment with local objectives. Number of items that assess individual objectives may be insufficient. Require attention to ensure that test results are aggregated and displayed appropriately.
Publisher's customized tests	Should be closely aligned to local objectives. Professionally constructed, pilot-tested, and reviewed for bias. Validity, reliability and item-difficulty data may be available. Scoring and reporting services available.	Require examination of appropriateness and quantity of items that assess local objectives. Test construction, scoring, and reporting services may be very expensive. May be limited to paper-and-pencil multiple-choice items.
Publisher's textbook tests	Probably closely aligned to instruction. Readily available. Inexpensive.	Require examination of alignment with local objectives. Information about validity, reliability, and item bias not readily available. Procedures for printing, scoring and reporting uniformly may have to be developed locally.
District locally developed tests	Especially appropriate for assessing local objectives. Local credibility may increase use. Small scale allows flexibility beyond paper-and-pencil measures.	Difficult, time-consuming process. Validity, reliability and item bias must be examined locally. Procedures for printing, scoring, and reporting must be developed locally.
Uniform procedures requiring written performance by students scored according to a uniform rating scale	Particularly effective for assessing writing skills. Results are useful instructionally. Technical assistance available through ESCs.	Raters must be trained. Rating process is time-consuming.
Uniform procedures requiring other (nonwritten) performance by students scored according to a uniform rating scale	Especially appropriate for assessing some objectives.	Difficult to develop, standardized prompts must be written, uniform rating scales developed, and raters trained. Validity, reliability and bias may have to be examined locally.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 4

Alignment: A Definition

Alignment simply means matching. In educational context, it has numerous applications and may refer to a match between:

- local objectives and assessment,
- local objectives and teaching-learning activities,
- teaching-learning activities and assessment,
- state goals and local objectives, or
- objectives at one grade or school level and objectives at another level.

In the broadest sense, alignment means coordination among all the elements of objectives/curriculum, instruction, and assessment. A well-functioning alignment provides smooth, cyclical transitions from planning to instruction to assessment, remediation, and enrichment to evaluation and then back to planning. Everything works together. The elements are coordinated between elementary and secondary schools, as well as between grade levels in each.

Many factors affect the extent of alignment within a district, but successful alignment is most evident in districts where:

- communication among teachers and administrators at all levels is open and functional, that is, communication channels are purposefully used to support or increase alignment;
- educational goals and objectives are written, coordinated across grade levels, and well known to, and supported by, teachers and administrators at all levels.

*Chapter adapted from Alaska *Assessment Handbook.*

Of course, just because a district has goals and objectives in writing does not necessarily mean that teachers are teaching those things, nor that tests are measuring what teachers are teaching. Administrators and teachers must monitor alignment to ensure that it is occurring.

The primary interest here is the integration of assessment procedures with objectives and instruction. Because these three areas are so intertwined, this chapter discusses the overall issue of alignment.

Steps toward Improving Alignment

Once a district has decided to improve alignment, how does it proceed? District staff must recognize that alignment is a dynamic process involving many continuously changing factors. A common sense approach is to think of alignment as a continuum. It can always be improved and works better in some circumstances than others.

Given that basic uderstanding, local staff may follow these steps toward improving alignment:

1. Review state goals and district learning objectives, looking for consistency among grade levels, an ordered progression from one level to the next, and appropriate tasks for each grade level.

2. Write/review/revise objectives to match what the goals indicate will be achieved.

3. Analyze the objectives to determine precisely what should be introduced, reinforced, reviewed, or mastered at each grade level.

4. Ensure that all tasks identified through Step 3 can be covered by instruction and that performance on these tasks can be assessed through testing or other viable procedures (e.g., classroom observation). Consider writing some sample test items to identify objectives that are too broad.

5. Devise a plan for the selection of textbooks and other materials to support knowledge and skill development in specified areas.

6. Outline a long-range plan for assessment that satisfies the district's specific needs.

Practical Tip

Alignment is most beneficial when all elementary and secondary grade levels are involved. That way, programs and expectations of students are more coordinated and consistent from level to level. Alignment should involve teachers from all elementary and secondary schools that students are scheduled to attend—even if that means working with teachers from another district.

7. Establish procedures for test selection and/or assessment development. The procedures should emphasize criteria that foster alignment (i.e., a match between test content, objectives, and teaching-learning activities). Assign responsibilities for assessment selection, development, and review.

8. Review current instructional plans (including Learning Assessment Plans) and assessment data to identify which goals and objectives are being addressed well, poorly, or not at all.

9. Recommend new instructional strategies that support alignment (e.g., integrating science and math instruction to develop students' math skills through science problems such as farm management).

10. Survey teachers to determine staff development needs. Do teachers need more information and skills related to assessment? Teaching across learning areas? Writing measurable objectives? Designing instruction for selected objectives?

11. Design staff development based on the findings of Step 10.

12. Provide forums for representatives of various groups (teachers, administrators, and community members) to share their perceptions about objectives, instruction, and assessment at each grade level.

13. Review local expectations for student achievement. Are they still appropriate after assessment data are reviewed and objectives or instructions are revised? Or should the expectations be raised or lowered?

14. Ensure that the assessment program adequate ly measures the achievement of expectations, ensure that the instructional program gives stu dents the skills and knowledge needed to meet expectations.

15. Review local remedial and enrichment programs. Are placement procedures consistent across the district? Are those procedures directly related to instructional and testing programs?

Practical Tip

Many districts 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 may have already tried the procedures, know what works and what doesn't. Other districts can adopt the good and correct the bad—with much less effort than starting from scratch.

16. Make alignment an inherent part of future planning and of high priority. Consider all perspectives (those of teachers, district curriculum coordinators, and other content or measurement specialists).

Districts cannot expect to go through these steps for all learning areas in a year. But they should develop plans for systematically and strategically improving this alignment of curriculum, instruction, and assessment even if it takes several years. The advantages for students and educators alike are well worth the effort.

The appendix includes a checklist which can be used in several alternative ways:

—as a survey to give to teachers and others early in the alignment process to help identify local needs,

—as a worksheet for the local planning group to use in examining the current status of alignment locally, and/or

—as a survey or worksheet to use after several years have passed to help determine what has been accomplished and to identify current needs.

Districts may want to adapt the checklist to fit local conditions and to consider other types of alignment also. For example, they may want to examine the compatibility between local educational philosophy and objectives/instruction, objectives and instructional processes, or assessment purposes and procedures.

Special Issues to Be Resolved

Regardless of how successfully or smoothly a district may handle its alignment efforts, certain problems can arise. Two common problems and possible solutions are described below.

- Problem: Maintaining the autonomy of each level in the education system. What if elementary and secondary teachers see themselves as very different in philosophy or approach? Do secondary staff have the right to dictate what should be taught or emphasized? Consistency of curriculum across levels is critical, and when one level dictates to or directs another, the spirit of cooperation necessary for efficient alignment may be lost.

  Solution: Provide a forum for discussion among representatives of all levels. No single level should take the lead in setting educational objectives or priorities. Educational objectives and curriculum should reflect the emphases that educators from all levels view as critical.

- Problem: Ensuring that testing plays a realistic role. Since some important educational outcomes are not measured through tests, good alignment does not demand that everything in the curriculum be tested formally. Misunderstanding of this can lead to overtesting which, in turn, leads to other problems such as scheduling conflicts, student anxiety, and staff resistance.

  Solution: Alignment policies should not overemphasize assessment; the match between objectives and instruction is just as important as the match between objectives and assessment. Clearly state how objectives should be selected for assessment. Emphasize other valid ways of assessing students' competence, including careful classroom observation.

Practical Tip

Do top administrators understand and support alignment efforts? A lot of work can be wasted if alignment efforts don't have consistent support.

A Summary of Advantages

The potential advantages of alignment may already be apparent. They include:

- Improved communication,
- Better problem solving through coordinated effort,
- Smoother transition for students (between grades and between school levels),
- Improved instruction (through consistency), and
- Increased satisfaction and productivity as a result of a focused, consistent effort.

These advantages suggest alignment's potential impact on students. When objectives, instruction and assessment are aligned, a school's mission is more likely to be accomplished. Since, presumably, that mission is based on student needs, students are the prime beneficiaries of the alignment process.

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 its 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.

Discusses how testing and evaluation should fit into classroom instruction, including how tests can help 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 a national standardized test to its 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.

Thoroughly reviews 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 MCTs 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 following checklist is one approach to determining alignment. Use the checklist to judge your district's progress toward improving 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 student learning objectives have been established and put in writing	0	1	2	3	4
2. The objectives are coordinated across school and grade levels,* with an ordered progression from one level to the next	0	1	2	3	4
3. The objectives match what the State goals indicate will be achieved	0	1	2	3	4
4. The district objectives are known and supported by teachers 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 support the skills and knowledge 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

Assessment

9. Assessment consistent with district objectives occurs on a regular basis	0	1	2	3	4
10. Assessment is matched to course content and materials	0	1	2	3	4
11. Assessment is matched to classroom instruction	0	1	2	3	4
12. Assessment results are used to evaluate which goals and objectives are being achieved and which are not	0	1	2	3	4

*Even in non unit districts, the objectives are coordinated between elementary and high school districts.

Communication

13. Open and functional communication exists among educators at all levels within the district*	0	1	2	3	4
14. Assessment results are clearly communicated to the public	0	1	2	3	4
15. Forums allow interested groups (e.g, parents and teachers) to share ideas about objectives, instruction and assessment	0	1	2	3	4
16. 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	16		

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 throughout the alignment effort.

Next, circle the appropriate numbers for all those criteria that cannot be met, even with special effort.

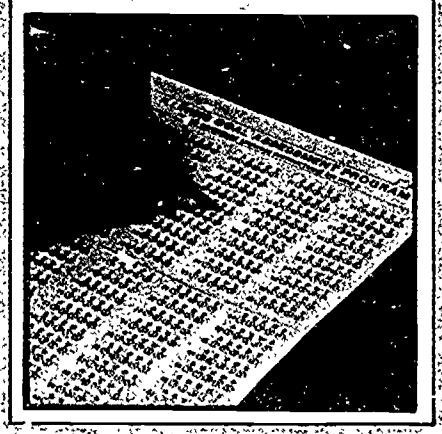
1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16		

You may not have to circle any numbers, but, more likely, one or two important components of the alignment effort 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.

*In non-unit districts, open and functional communication also exists between elementary and high school district educators.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 5

Purposes of Standardized Norm-Referenced Tests

For most school districts in Illinois—and in the United States—standardized norm-referenced tests (NRTs) are an important part of district assessment programs. These commercially available tests were prepared by measurement experts. When uniform procedures are used to administer and score the tests, local student performance can be compared to national norm groups who were tested using the same instrument.

This ability to compare test results makes standardized tests more useful for some purposes than other types of test. In situations where students or student groups must be ranked (e.g., for some accountability and evaluation decisions or for selection of students for special programs), a well chosen standardized NRT should be seriously considered. Standardized tests are appropriate in accountability and evaluation when the question of interest is "How are our students doing compared to similar students throughout the nation?" When the question is "Are our students learning the skills and obtaining the knowledge we say we are teaching them?", standardized tests are less appropriate.

Most districts use standardized test batteries which cover a broad range of basic skills content, although standardized diagnostic and single-subject tests with associated norms are available. Many districts, particularly larger districts, also use an aptitude test which was normed on the same sample as the publisher's achievement battery. This allows a two-way comparison of their students' performance—externally (compared to similar students nationwide) and internally (compared to their own expected achievement as determined by the aptitude measure). In some districts, this two-fold comparison is an efficient use of testing resources for gathering a maximum amount of information.

*Chapter adapted from Alaska *Assessment Handbook.*

To be valid and beneficial, standardized achievement tests must measure the appropriate skills and knowledge that, generally, are described in local learning objectives. The previous chapter included suggestions for improving that alignment. Guidelines for selecting standardized tests that are most useful locally are provided in this chapter.

Criteria for Selecting Standardized Tests

District planning groups may use many different criteria to guide their selection of standardized tests. Those criteria are suggested and described here. The Chapter appendix includes worksheets that local groups can use, perhaps with adaptations to fit local conditions, in the selection process. The criteria are categorized into three types of selection considerations. alignment, technical and practical.

The first worksheet "Content Validity of Standardized Tests" is for examining the alignment between local objectives and a particular standardized test, as discussed in the next section. Information from that worksheet will be used to complete the first section of the second worksheet "Rating Sheet for Standardized Tests."

Alignment Considerations

Alignment criteria are critical in test selection. Standardized tests that are not closely aligned to local objectives should be eliminated from further consideration.

One important question—which should be answered "yes" if a test series is to receive further consideration—is whether a majority of the test items match local learning objectives. If not, it would be a very inefficient instrument to use, it would provide the comparative information desired from an NRT, but would be of little help in curriculum evaluation.

The next consideration is the proportion of objectives which are measured by test items. Local planners may decide that a test is inefficient if it assesses only a small percentage of local objectives—unless the test measures important objectives very well and can be used for multiple purposes. This criterion may also be an indicator of the level of specificity of local objectives. If the percentage is extremely small for all the test series considered, local objectives may be too specific and need to be broadened.

If an NRT is used to assess individual objectives for accountability or diagnostic purposes, at least 3 5 test items per objective (and the more, the better) are recommended. With fewer items, a student could guess the correct answer and appear to have "mastered" the objective.

The final alignment consideration is whether test items and district objectives receive comparable emphasis. If a district uses the worksheet in the appendix to determine content validity, the more important local objectives should appear in the appropriate column more often than less important objectives.

Technical Considerations

Good reliability, i.e., consistency of test scores, is a necessary, but certainly not sufficient, characteristic of a test. How much is "good"? A test whose reliability is lower than .80 (the theoretical limit is 1.0) should be viewed with skepticism, .90 or better is commonly achieved by national norm referenced basic skills tests.

The higher the reliability, the more confidence a district can have in test scores. This philosophy suggests that high reliability is needed in those tests which are used for making important decisions and is especially important when reporting individual student scores rather than group scores.

Subtest reliabilities are most often lower than total test reliabilities because reliability depends a great deal on the number of items included in a score. Therefore, not every subtest reliability will be in the .90 range. Still, a test that has lots of low subtest reliabilities should be avoided.

Expert opinions about standardized tests can be found in a number of sources. (See the Halpern article referenced at the end of this chapter.) Districts should still study the technical manuals for each series and examine test items, but the reviews can save time in narrowing the field of potential tests.

Tests differ in the groups of students used to establish norms. Some districts, especially those in very urban or rural areas, should review information about tests' norming samples. For example, rural districts may prefer not to use tests normed on students who are mostly from large urban districts and instead choose tests whose norming samples include a substantial representation of students from rural districts.

Practical Considerations

Practical considerations—including questions about test format, administration, scoring, costs and publisher's services—may distinguish between an acceptable and an unacceptable test series. 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 that use the same test may help answer questions about practicality.

Advantages and Disadvantages of Commercial Tests

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

Advantages include:

- High technical quality,
- Norms to allow comparisons with external groups,
- Free consulting from publishers' representatives,
- No developmental costs.

Disadvantages include:

- No reason to expect good match between what is tested and what a district emphasizes in its teaching,
- Often too few items per objective to allow the test to be used for diagnostic or certain accountability and evaluation decisions (e.g., determining the proportion of students who meet an objective),
- May be prohibitively expensive—or even impossible—to get exactly the kind of reports needed,
- Recurring annual costs for materials and (perhaps) scoring.

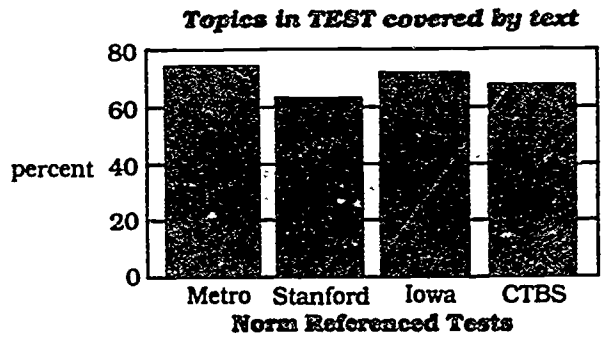
While the last disadvantage is unavoidable, the impact of the first three can be reduced by careful selection and planning.

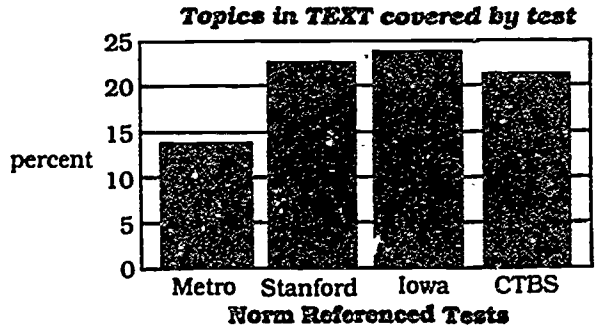
Choosing among Tests: Does It Really Matter?

Content emphasis differs among the various standardized achievement tests. The two charts below emphasize this point graphically. They are based on data from a study conducted by the Institute for Research on Teaching (IRT) at Michigan State University (Freeman, et al., 1983). That study compared the content coverage of commonly used fourth grade math textbooks and standardized tests.

The charts show the comparisons between several tests and Scott-Foresman's *Mathematics Around Us*. The top chart shows that for even the best-matched test (the Metropolitan), 25 percent of the test items are not covered in the text. The test with the best match on coverage (the Iowa—see bottom chart) still measures less than 25 percent of the topics covered in the textbook.

The graphs show that since no standardized test perfectly measures the curriculum, there will always be tradeoffs. The matching activity described in the next section can help districts decide what tradeoffs to accept.





References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing.* (This document has been distributed to all Illinois school districts.)

This publication contains standards which can be used to evaluate published tests.

Freeman, D. J.; Kuhs, T. M.; Porter, A. C.; Floden, R. E.; Schmidt, W. H.; and Schwille, J. R. (1983). Do Textbooks and Tests Define a National Curriculum in Elementary School Mathematics? *The Elementary School Journal, 83 (5),* 501-513.

Reports the results of a study of the relationship between testing and the curriculum.

Gronlund, Norman E. (1985). Test Selection, Administration, and Use. Chapter 13 in *Measurement and Evaluation in Teaching.* MacMillan, 320-345.

Two sections of this chapter deal with "Obtaining Information about Published Tests" and "Selecting Appropriate Tests." A suggested test evaluation form is included.

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.

Includes a section on general principles to be used in choosing a test. Sources of test reviews are also discussed.

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* (Sample Form)

SUBJECT _____ GRADE _____ DATE _____

REVIEWER _____

TEST _____ FORM _____ LEVEL _____

Directions: In the first column, write the item numbers of the test you are reviewing. In the second column, write the code numbers of the local objectives which match the test items. Use as many sheets as necessary to cover all the items in the test. For each test item, consider these three statements:

- The item matches a district learning 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. Enter those codes in the last three columns.

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

NOTE: More extensive directions for completing this form are on the next page.

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

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

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

*Chart adapted from Alaska *Assessment Handbook.*

Directions for Completing Sample Form

1. Number the local objectives in the learning area under review. Identify each 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 test objectives are in line with local objectives.

3. In the first column, write the item numbers of the test you are reviewing. Read each item and decide if it measures one or more local objectives. Do not rely on the test publisher's description or item classification chart. Enter the objective number(s) in the second column. (When reviewing many levels of a test, consider sampling; randomly select 40-50 items at a minimum of three levels.)

4. For each item that matches an objective, consider the following statement:

   "The item matches a local learning objective closely."

   Use the following scale to indicate how much you agree with the statement:

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

   Enter the appropriate code number in the third column.

5. Rate the quality of the item, indicating whether you agree or disagree with the following statement:

   "The item is of high quality."

   Using the same 5-point scale, enter the appropriate code in the fourth column.

6. Rate each item's grade-level appropriateness. Again using the 5-point scale, indicate in the fifth column the extent of your agreement with the following statement:

   "The item is at an appropriate level of difficulty for the chosen grade."

7. Determine how many test items appropriately measure local objectives. If fewer than half the items match objectives, the test does not fit your curriculum very well. (You may decide to do this after Step 3, thus eliminating tests that match poorly before you do the work required in succeeding steps.)

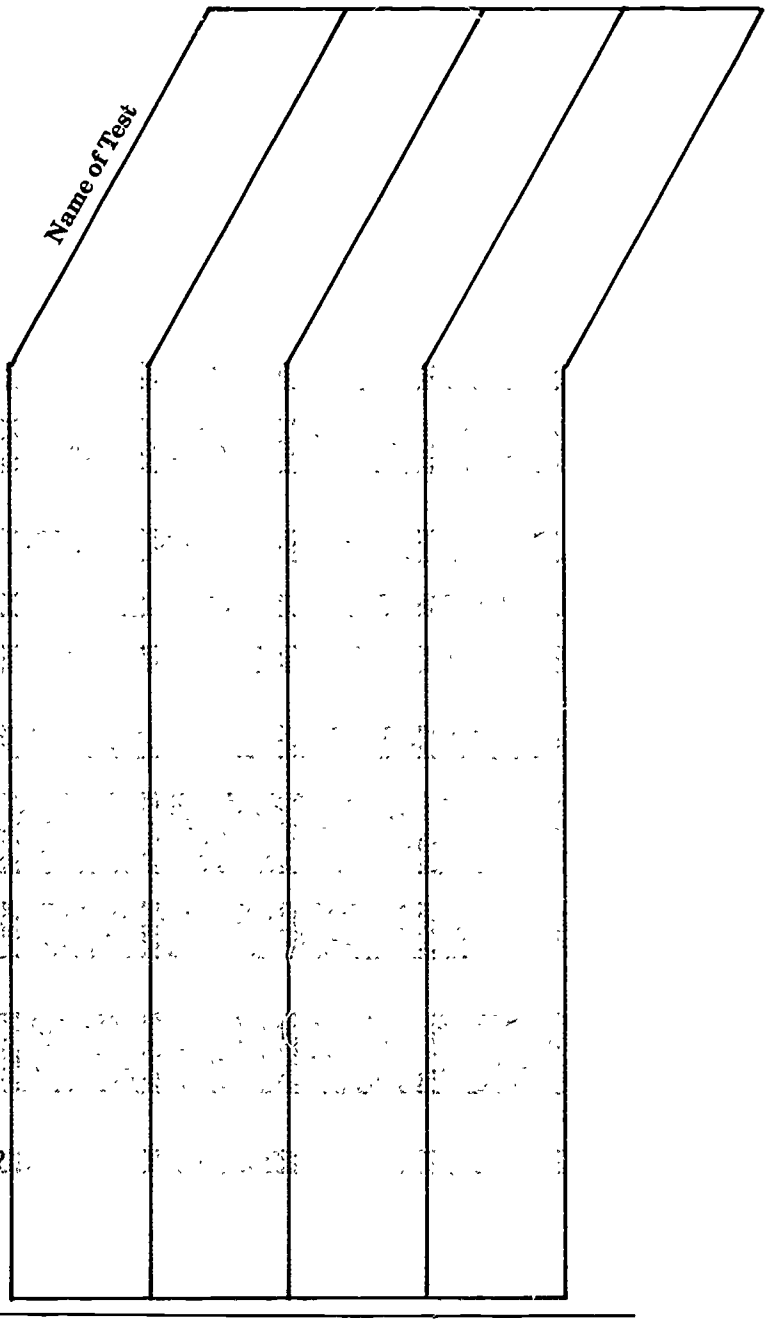
Rating Sheet for Standardized Tests*

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. Tests should not receive credit for missing information. Most test publishers know that tests cannot be evaluated without adequate technical information, missing information reflects negatively on the test.)

The previous worksheet "Content Validity of Standardized Tests" can be used to answer the alignment questions. The first alignment item—the match between test items and district objectives—should be answered positively before a test is considered further. Beyond that, the test with the highest rating may be the one a district should choose although individual circumstances may cause a district to include additional criteria or to weight them differently.



ALIGNMENT CONSIDERATIONS

1. Test items match district objectives?
2. High percentage of district objectives measured?
3. Multiple items per objective (at least 3 if 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 opinions held by expert reviewers?
7. Normed recently?
8. Norming sample appropriately representative?
9. Empirical norm dates match district's testing schedule?
10. Item quality acceptable?

PRACTICAL CONSIDERATIONS

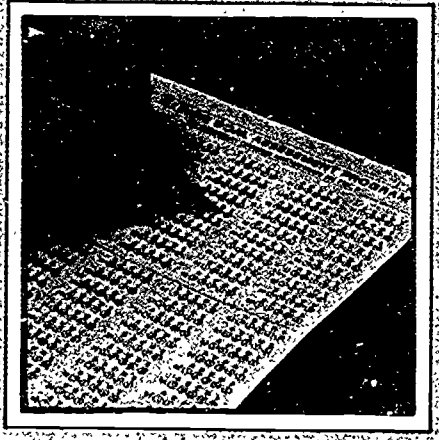
11. Format (e.g., number of items per page, print size, response mode) appropriate for grade level of students?
12. Items free of sex, cultural and ethnic bias?
13. Easy for teachers to administer and, if necessary, score? Not too time consuming?
14. Cost for consumables, scoring, reporting, and other services within budgetary limitations?
15. Needed score reports (e.g., for clusters of items that assess local objectives) available? Consulting and other assistance?
16. Adequate coverage (enough test levels for grades scheduled for testing)?
17. Alternate forms available if district wants them?
18. Related tests available if district wants them (e.g., co-normed aptitude measure or achievement tests for other learning areas)?

TEST TOTALS

*Chart adapted from Alaska *Assessment Handbook.*

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 6

- Item Banking
- Writing Assessment Items
- Constructing Tests from Items
- Developing Performance and Other Alternative Assessment Procedures

Many local planning groups may decide to construct at least some local tests, assessment items, or other assessment procedures to use in conjunction with other approaches such as commercial tests. Although local construction is a difficult, time-consuming process that demands rigorous attention to quality, many resources are available to help guide it. And, the rewards for having good assessment procedures that are particularly appropriate or valid for local objectives can be worth the efforts.

Local staff who decide to construct assessment procedures can use several different sources of items. They can purchase commercial item banks, use an item bank developed by the Illinois State Board of Education (ISBE) and available through local Educational Service Centers (ESCs), and write their own items.

Item Banking

Item banks are collections of assessment items. Local staff can review the items and select those which they want to use. Some item banks are computerized; others are simply stored on paper. Generally, the items have been pilot-tested with a sample of students, and several types of information are available for each item. For example, items in the bank developed by ISBE indicate the learning area, goal, and knowledge/skill assessed. The items also include data from samples of Illinois students who took the items during a spring 1987 pilot assessment. The data include the grade level of the students who took each item, the proportion of students who selected each response alternative, and other variables.

Item banks can be very useful sources of items. Most contain hundreds, perhaps thousands, of items. Many items have been carefully constructed, reviewed, and pilot-tested. Data about item difficulty can help guide decisions about whether or how to use an item. Of course, districts must review the items carefully to determine the appropriateness for measuring local objectives.

A few districts may want to build their own banks of assessment items. They may begin with a collection of items from various sources and gradually add other items they obtain or write. Having access to a large number of items that have been used locally can be very advantageous in future assessments. Alternative forms of good tests can be constructed relatively easily.

Writing Assessment Items

District personnel who write their own assessment items must devote considerable resources, including training time, to that task. (Educational Service Center staff may be able to assist with that staff development.) Districts need to allocate at least the equivalent of several weeks of work to the writing, piloting, and revision of items in each selected learning area.

Many guidelines are available to help local staff write assessment items. One source *Improving Your Test Questions* by John C. Ory of the University of Illinois is included in the general appendix. Other resources are listed at the end of this chapter.

After writing items, staff should edit them carefully to control their quality. (Checklists for reviewing items—locally constructed as well as items obtained from elsewhere—are shown in *Improving Your Test Questions.)*

Finally, staff should pilot-test the items by administering them to a number of students, at least 10 students per item. While doing this, they should ask several questions about each item.

- Does the item appear to assess the intended knowledge/skill?
- Do students clearly understand the instructions and the items? (Talk to some students about why they responded as they did.)
- In comparison with other information such as teacher judgment, do those students who might be expected to respond correctly (or incorrectly) do so?
- Are any distractors (incorrect response options) confusing or too blatantly incorrect?
- Does the distribution of responses across different sex and ethnic groups indicate that the item is free of bias?
- Is the item at an appropriate level of difficulty?

Constructing Tests from Items

Local staff groups who write or acquire a large number of test items have taken a major step toward constructing their own local tests. Test construction, however, involves more than simply assembling the items into booklets. To construct tests, staff will need to perform the following tasks.

Pilot-test the items. This task, discussed in the previous section, will be necessary only if the items have not already been pilot-tested with students who are similar to local students.

Identify the test purpose. Several types of decisions about test construction will be influenced by the intended use of the test results. For example, the level of difficulty of items in a test for assessing student mastery should be different than in an achievement test. Most items in a mastery test should probably be at a level (e.g, 80%-85%) which indicates that students have learned the content. An achievement test should include items at a wide range of difficulty levels. Test validity must be estimated in relation to test purpose. As will be discussed more extensively later (in Chapter 7), any test will be more valid for some purposes than others.

Develop test specifications. This stage, sometimes referred to as developing test "blueprints," involves making decisions about the composition of a test. The test construction committee will specify how the test items should be distributed across one or more factors.

Two factors the committee might particularly consider including in the test specifications are content and item difficulty level. Developing content specifications should include identifying categories of information to be covered and deciding what proportion of the items should be devoted to each. Such categories might represent headings in a curriculum outline, knowledge/skill statements, or local objectives. Staff might want to include an equal number of items for each category, or they may decide that some topics receive more instructional emphasis than others and should receive comparable emphasis on the test.

As mentioned previously, the difficulty level of the items should probably vary according to the intended uses of the test results. For example, a test for identifying the proportion of students who have achieved an objective or mastered a particular content area should consist primarily of items which indicate mastery. A test for sorting students according to their achievement in a particular area should contain items that are at a variety of difficulty levels.

Staff may also identify other variables to include in the test specifications. They may want a test to include items that are at several different cognitive levels. For example, they may decide that most items should assess factual knowledge, but that others should assess whether students can apply that knowledge. Staff may want to develop test specifications for multiple types of content categorization. For example, in world geography they might examine item distribution across knowledge/skills and across continents. Specifications for history, literature, or fine arts tests might include identifying item distributions across major historical periods or cultures.

A sample matrix for specifying the contents of a test for State Learning Goal #3 in mathematics is enclosed in the chapter appendix.

Assemble the test. This stage will include three major activities: 1) selecting items, 2) arranging them into test booklets, and 3) developing instructions for standardized administration of the test. When selecting items, staff should, of course, refer to the test specifications. However, staff should also review each item carefully. They should consider whether it accurately reflects local objectives/instruction and whether it assesses knowledge or skills they think are important. Local staff should also examine the information that is available from previous administrations of the item. What pro portion of students selected the correct response? Does the distribution of responses to incorrect alternatives suggest that the item is poor because one or two alteratives were so obviously incorrect that only a few students (less than 10%) chose them? Do the item statistics (for example, the point biserial correlation coefficient) indicate that the item functions accept ably well?

Staff will probably decide to arrange the test items by level of difficulty (from easy to hard) or by content area. Arrangement by difficulty level increases the opportunity for students to show what they know by allowing them to answer the easy items before time runs out or before they become discouraged by encountering too many difficult items. On the other hand, arrangement by topic may make it easier to summarize the data and will permit coverage of major topics first and less crucial topics at the end of the test. Other alternatives include arranging items by type (for example, keeping those requiring special instructions or graphics together) or by other content categories.

Standard administration instructions are important to ensure that the tests are given uniformly to all students. The instructions might indicate, for example, directions to be given to students (Unclear directions may unfairly reduce students' scores.), the resources that students may use during the test (e.g., books, calculators), whether guessing is allowed or encouraged, and the amount of time allotted for the test.

Field-test and revise the instrument. Before tests are used widely, they should be given to a small but representative sample of students (10 or more, depending on local circumstances) using procedures similar to those described previously for piloting items. Each time the test is revised, the new version should be field-tested. Following these procedures (with items and with tests) should limit the amount of time and resources that are lost because tests do not perform as expected.

Review the test for validity and lack of bias. This stage may occur before or after field-testing. Regardless, major changes in the instrument will require additional reviews or field-testing. The reviews should be conducted by panels which include teachers and others who are knowledgeable about the content area assessed or with bias review procedures. The panels should be independent of the test construction committees; people who review tests for validity or nondiscrimination should not have been closely involved in test development. Validity and bias review will be discussed more extensively in the next chapter.

Developing Performance and Other Alternative Assessment Procedures

Most local objectives can be assessed with the multiple choice (or matching, true-false, or fill-in-the blank) items that are common to many published and locally constructed tests. However, some objectives, such as those requiring students to "demonstrate" various behaviors or those related to the State goals listed below, can be assessed most effectively using alternative procedures such as rating student writing or other performance. Those state goals which may require alternative procedures are listed below.

As a result of their schooling, students will be able to:

- write standard English in a grammatical, well-organized, and coherent manner for a variety of purposes (language arts, state goal #3);
- use spoken language effectively in formal and informal situations to communicate ideas and information and to ask and answer questions (language arts, state goal #4);
- demonstrate the basic skills necessary to participate in the creation and/or performance of one of the arts (fine arts, state goal #3);
- demonstrate basic skills and physical fitness necessary to participate in a variety of conditioning exercises or leisure activities such as sports and dance (physical development and health, state goal #4);
- perform a variety of complex motor activities (physical development and health, state goal #6); and
- demonstrate a variety of basic life-saving activities (physical development and health, state goal #7).

Alternative assessment procedures can also sometimes provide information that supplements and enriches data from other types of assessment. For example, a student might be able to answer questions about the scientific process accurately, but still be unable to perform a laboratory experiment properly. Observation of the latter would provide useful information.

Currently, few performance or other alternative assessment procedures are available commercially, except in writing assessment. However, several sources of information about alternative procedures are available. As mentioned in Chapter 3, procedures for assessing student writing have been used widely and tested extensively in Illinois schools. Information about those procedures is available through ESCs. In addition, suggestions for using and writing essay, problem-solving, and performance items are included in *Improving Your Test Questions* which is in the general appendix. Useful guidelines may also be found in Gronlund (1985, Chapter 15), Popham (1981, Chapter 13) and Roid and Haladyna (1982).

Alternative assessment procedures might include:

- Assessing student performance by observing it directly (for example, observing a student performing a motor skill or participating in a dance or debate);
- Focusing on processes used in performance or in product development (e.g., making preliminary sketches of a drawing, interpreting music notation, or deciding how to use tools such as the body or props in a dramatic performance);
- Assembling and rating portfolios or other artifacts of student work such as:
  - —visual artworks (photographs, paintings, pottery),
  - —audiotapes of music performances,
  - —dance choreography,
  - —videotapes of dramatic performances, or
  - —journals or sketchbooks;
- Assessing the performance of students individually or as members of a group such as a sports team or band;
- Having student performance assessed by various persons: the individual students themselves, peers, teachers, content experts such as artists, or audience members;
- Using external stimuli (for example, reproductions of visual art, audiotapes of musical performances, or videotapes of dance or dramatic performances) to assess various kinds of knowledge with multiple-choice tests.

When adopting performance or other alternative assessment procedures, local planning groups *must* develop standardized instructions for eliciting the relevant student behavior, as well as criteria and systematic methods for assessing that behavior. Teachers should be trained in the process. The information must be thoroughly documented and maintained in a centrally located file.

References

Gronlund, N.E. (1985). Constructing Classroom Tests. Part Two of *Measurement and Evaluation in Teaching*. New York: MacMillan Publishing Company, Inc.
Contains guidelines for writing items and constructing tests. Part Four discusses alternatives to paper-and-pencil tests.

Illinois State Board of Education. (1987) *An Assessment Model for Mathematical Problem Solving.*
Describes a problem solving assessment model that reflects key curriculum changes. May be very useful in the development of local mathematics assessment items.

Illinois State Board of Education. *Write on Illinois!*
Two volumes of this document are available. The second is an updated and expanded version of the first (Two copies of the second volume have been distributed by Educational Service Centers to each Illinois school district.) Both volumes are users' guides for scoring student essays. They include information about the Illinois writing assessment program and might be useful to districts in the development of procedures for local writing assessment.

Lockwood, R.E. and Hess, A.C. (1984). *Test Construction. A Writer's Guide.* Montgomery, AL. Alabama State Department of Education.
In addition to directions for writing items, it includes helpful suggestions for establishing test and item specifications and for analyzing test data.

Ory, J. *Improving Your Test Questions.* Urbana-Champaign, IL, University of Illinois, Office of Instructional and Management Services.
This booklet, which is included in the general appendix, contains very clear, practical guidelines for writing test items.

Popham, W.M. (1981). Creating Educational Tests. Part III of *Modern Educational Measurement.* Englewood Cliffs, N.J.: Prentice-Hall, Inc.
Includes guidelines for writing and revising test items, constructing tests and using observations and ratings.

Roid, G.M. and Haladyna, T.M. (1982). *A Technology for Test-Item Writing.* New York. Academic Press.
This book focuses primarily on item writing. It includes suggestions for constructing and using rating scales and checklists (Chapter 5), for developing higher level test items (Chapter 10), and for reviewing test items (Section IV).

Urbana School District 116. (1986). *Student Assessment System for Urbana School District 116.* (Available at Educational Service Centers.)
Includes an exhibit document that contains competency based math tests – which should *not* be used without permission from the Urbana district.

Appendix: Chapter 6
Sample Test Specification Matrix
Fundamental Learning Area—Mathematics

Areas of Focus - Test Design

Mathematics

State Goal for Learning 3

As a result of their schooling, students will be able to make and use measurements, including those of area and volume.

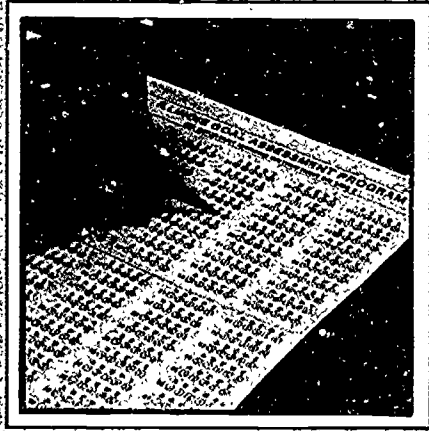
GENERAL KNOWLEDGE/SKILLS RELATED TO GOAL 3		Recall	Computation	Understanding	Problem Solving
The following knowledge, processes, and skills are related to this State Goal for Learning:					
A Measurement in various contexts using appropriate units	A				
B Estimation of measurements.	B				
C Relating lengths, areas, and volumes in common geometric figures	C				
D Conversion of units within one system and from one system to another.	D				
E Application of selected measurement systems, instruments, and techniques.	E				

Mathematics Abilities

1. RECALL: the ability to recall and recognize facts, definitions and symbols quickly. Perception is the primary mental act used.
2. COMPUTATION: the ability to perform computations, procedures, and complex counting where the operations are indicated.
3. UNDERSTANDING: the ability to understand concepts, facts, and processes. The mental operations of analysis and synthesis are used to make comparisons and evaluative judgments.
4. PROBLEM SOLVING: the ability to solve complex word problems. Several of the following operations must be involved: interpretation of the question, identification of the relevant data from the given information, decisions about which operations need to be performed on the data, correct performance of the operations, and interpretations of the results.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 7

- Maximizing Validity and Reliability
- Avoiding Bias
- Using Multiple Approaches
- Assessing Different Cognitive Levels
- Using Assessment Procedures for Multiple Purposes
- Developing an Assessment Schedule

When developing assessment systems, local planning teams may ask questions such as.

How do we ensure that assessment is as valid and reliable as possible?

How do we avoid biasing the assessment against certain groups of students?

Can we adopt a commercial test, a locally developed test, or some other assessment procedure that will meet most assessment needs?

Are we focusing too much on assessing lower-level cognitive skills?

How can assessment become more efficient?

What should we consider when we develop an assessment schedule?

Local staff will return to some of these questions repeatedly. For example, validity, reliability, and bias are important when selecting commercial tests (or items), developing local items/tests, planning how to use them, and incorporating them into a comprehensive assessment system. Later, staff should return to these topics when administering the assessment, interpreting the results, and making consequent decisions.

Maximizing Validity and Reliability

Districts are required to provide assurances regarding the validity and reliability of each assessment procedure included in local Learning Assessment Plans (LAP). Both are discussed here, but districts should also consult other resources such as those listed at the end of this chapter.

Validity

Validity is the extent to which a test or other assessment procedure is capable of producing information that warrants a particular type of interpretation. For example, does a test adequately represent a particular objective or set of objectives? Can data from the test provide a reasonable estimate of student accomplishment of the objective(s)? Validity must always be estimated in the context of a particular intended use of data from an assessment procedure. Most assessment procedures are more valid for some purposes than others. For example, a test may be excellent for estimating the overall status of a district's students on a specific set of learning objectives in mathematics. However, that same test might be much less useful for diagnosing the needs of individual students and completely inappropriate for evaluating the effects of a mathematics program.

Districts may accumulate evidence of validity using at least three different strategies. 1) obtaining evidence about the validity of published tests from test manuals or by writing to the publishers, 2) asking a panel of teachers and other content experts to review an assessment procedure to estimate its validity, and 3) gathering empirical data to examine validity. Districts should attempt to accumulate multiple kinds of evidence of validity. Each should be associated with a particular intended use.

Local planning staff may decide to examine validity by asking local panels to review content validity. Such panels are most likely to review locally constructed items/tests, however, reviews will also help establish appropriateness of publishers' tests for designated local uses. The local panels should include teachers and other content experts (such as curriculum specialists) who are thoroughly familiar with local objectives and instruction. Assuming that a panel is examining the validity of an assessment procedure for estimating whether students have attained local objectives, the panel's major task is to compare the assessment procedure to local objectives.

A panel that examines validity might ask questions such as:

> Do the items in the assessment procedure adequately represent the objectives that are purportedly measured?
>
> Are there at least 3-5 items for each objective? (More may be required for certain test uses such as diagnosing the instructional needs of individual students.)
>
> Are some knowledges/skills neglected? Overrepresented?
>
> Are some items irrelevant?
>
> Are the items at an appropriate level of detail for the intended use?

Ideally, districts should assemble local panels every few years to review the continuing validity of assessment procedures. Objectives and instruction are likely to change over time. Validity panels which systematically review local assessment procedures can help ensure that the procedures remain valid.

Planning groups which decide to gather their own empirical data to examine validity can use references such as Gronlund (1985) or Popham (1981) for guidance.

Reliability

Reliability refers to the consistency or stability of assessment results. Several different types of reliability have been identified. One of the most common refers to consistency across time. It asks, for example, the extent to which assessment results would have been the same if a test had been administered a few weeks earlier or later. Or, did factors such as student guessing, fatigue, or motivation strongly affect the results? Unreliable assessment results are not, of course, worthy of use in educational decision making of any kind.

Local staff can obtain estimates of reliability from test publishers or develop their own estimates. For the latter, staff can refer to several resources such as Gronlund (1985) or others listed at the end of this chapter. They can also obtain computer programs that estimate reliability. Regardless of the source, local staff should carefully examine whether the method used to compute reliability provides evidence of the type(s) of consistency most important for their purposes.

Gronlund (1985) describes several different methods of estimating reliability, as shown in Table 7-1.

Table 7.1

Methods of Estimating Reliability*

Method	Type of Reliability Measure	Procedure
Test-retest	Measure of stability	Give the same test twice to the same group with any time interval between tests from several minutes to several years.
Equivalent forms	Measure of equivalence	Give two forms of the test to the same group in close succession.
(Test-retest with equivalent forms)	Measure of stability and equivalence	Give two forms of the test to the same group with increased time interval between forms.
Split-half	Measure of internal consistency	Give test once. Score two equivalent halves of test (e.g., odd items and even items); correct reliability coefficient to fit whole test by Spearman-Brown** formula.
Kuder-Richardson	Measure of internal consistency	Give test once. Score total test and apply Kuder-Richardson** formula.

*From Gronlund (1985), p. 89.
**For additional information about these formulas, see Gronlund (1985) or Popham (1981).

Another type of reliability, which is particularly important for performance assessment procedures, is inter-rater reliability. It is an estimate of the consistency of the ratings assigned by two or more different raters. Inter-rater reliability is estimated after multiple raters observe and assign ratings to the same students performing a particular task. The ratings are then compared. Inter-rater reliability can be improved through follow-up discussions in which raters talk about why they assigned particular ratings and attempt to develop common standards.

Several important points to remember about reliability are listed below:

- The time interval between test and retest is very important and often presents a dilemma. The interval should not be so brief that students remember answers from one test administration to the next. However, it should not be so lengthy that factors such as student learning or maturation are likely to cause actual changes.
- Reliability is a necessary but not sufficient condition for validity. Without reliability, validity is a moot question. What do assessment scores measure if they are not trustworthy? However, scores can be reliable but not valid; they can be very stable but still not assess the intended content.
- Reliability can be estimated only by using a statistical procedure. Unlike validity, it cannot be estimated judgmentally by a panel of experts. Also, reliability is applicable only to tests, not to individual assessment items.
- The split-half and Kuder-Richardson methods of estimating reliability may appear more feasible because they do not require a test to be administered twice. Indeed, they are very useful for examining the equivalence of two or more

test forms (including different forms constructed from item banks). However, they provide no information about stability across time.

The above discussion applies primarily to norm-referenced testing. The reliability of scores from locally constructed or other criterion-referenced tests can often be estimated using similar methods, but adaptations may be necessary. For example, Popham (1981) has suggested studying the consistency of decisions rather than scores. That is, how consistent—across time or test forms—are decisions such as the proportion of students who meet a set of objectives or cutoff score? Also, Popham points out that the smaller number of items on many local or criterion referenced tests may reduce the size of reliability coefficients, districts may need to decide that lower coefficients are acceptable.

Avoiding Bias

Districts must also provide assurances on their Learning Assessment Plans (LAPs) that they have taken steps to ensure that their assessment procedures "are nondiscriminatory in relation to race, sex, or national origin." Districts may also want to examine other potential types of bias such as discrimination against students with disabilities, from either urban or rural backgrounds, from low socioeconomic groups, or from specific religious groups.

Bias review procedures may include several different types of questions.

> Do any items perpetuate stereotypes about a particular race, gender, or national origin?
>
> Are various groups represented fairly?
>
> Are any items likely to be offensive to members of some groups?
>
> Have all groups had an equal opportunity to acquire the knowledge/skill that is assessed?
>
> Do answers to questions depend on knowledge that is not taught in school but that some groups are more likely than others to have acquired elsewhere?
>
> Is the language used likely to be reasonably familiar to all groups?

Most bias review procedures can be categorized as either judgmental or statistical. Judgmental review is usually conducted by a committee or panel which reads assessment items/procedures, asks questions such as those listed above, and identifies items which appear to be biased. Those items are then either revised or deleted. Statistical review involves examining data about the performance of various groups of students on the assessment items/procedures. A simple but useful type of statistical information is the proportion of students from each group who answered an item/procedure correctly (commonly known as the item difficulty level or "p" value). Items which appear to have been much more difficult for some groups than others should be reviewed judgmentally to determine whether the differences were caused by item bias.

Some bias review experts prefer either judgmental or statistical review procedures. However, a more thorough bias review can be conducted if both types of procedures are used interactively. Each has different strengths and weaknesses. Neither is sufficient alone. To counterbalance the weakenesses and capitalize on the strengths, both should be used.

Bias reviews should be conducted at several stages of the assessment cycle. Item writers should be sensitized to the types and sources of bias. Later, the items should be reviewed by others who are representative of various groups and knowledgeable about learning area content or statistical procedures. During the test selection process, the tests should be examined for bias. After tests have been administered, results should be reviewed statistically.

Districts might establish bias review procedures that specify. (a) the types of procedures that will be used, (b) the types of committees or panels that will be involved in the process, and (c) the stages of assessment at which bias review will be conducted. Guidelines for bias review are included in *Bias Issues in Test Development* (National Evaluation Systems, Inc., 1987), which was distributed to Illinois school districts previously. Before developing (or adopting) specific procedures, local planning groups should decide what they hope to accomplish through bias review:

- —to treat females and minority students fairly,
- —to prevent public controversy and lawsuits,
- —to avoid offending members of selected groups, and/or
- —to avoid perpetuating stereotypes.

Awareness of these purposes will help staff identify the types of questions to include in the procedures.

Using Multiple Approaches

Districts should use a variety of assessment procedures. As indicated in Chapter 3 and elsewhere, different types of procedures have different strengths and weaknesses. By using a variety of approaches, local staff can assemble a more comprehensive and accurate portrait of student learning. For example, a publisher's standardized shelf test may measure certain elements of a set of mathematics objectives but neglect others. Staff may decided to adopt locally developed items to complement a publisher's test. In another example, a district may have tests or items that assess students' understanding of the scientific concepts that are imbedded in a particular goal. However, those assessment procedures may not indicate whether students can perform related scientific tasks, e.g., design and conduct an experiment or analyze the composition of a chemical compound. Local planning staff may decide to assess these skills by observing students performing the tasks; planning staff would establish standardized performance situations as well as uniform criteria and procedures for rating student performance.

Assessing Different Cognitive Levels

When selecting or developing assessment procedures, staff should examine the level of student understanding that each assesses. For example, they may want to use the classification scheme (knowledge, comprehension, application, analysis, synthesis, and evaluation) presented in Bloom's *Taxonomy of Educational Objectives* (1956). Or, they may want to develop a more simple three-category system. Planning staff should make certain that they assess various cognitive levels and that local assessment procedures are not weighted too heavily toward the lower levels of knowledge and comprehension. To assist with this task, districts may develop checklists to indicate the level of understanding assessed by each item.

Using Assessment Procedures for Multiple Purposes

Generally, assessment procedures that are included in comprehensive assessment systems have been selected or developed primarily for specific purposes such as assessing students on a particular set of learning objectives. However, districts may be able to improve assessment efficiency by adopting procedures that can be used for more than one purpose. For example, some standardized publishers' shelf tests might be used to assess several different sets of objectives (e.g., in reading and mathematics), to report local and national comparisons to the public, or to examine the effectiveness of Chapter 1 programs. To the extent feasible, planning staff should attempt to identify assessment procedures that can be used for multiple purposes. However, the appropriateness of each procedure for its major intended use should always remain the top priority.

Developing an Assessment Schedule

Comprehensive assessment systems may include schedules which specify the grade levels and time of year each assessment procedure will be administered.

When deciding which grade levels to assess, local planners might consider factors/questions such as the following.

- P.L. 84-126 requires students in grades three, six, eight, and eleven to be assessed annually on local objectives in all six fundamental learning areas. (These requirements are being phased in under a schedule specified in the Rules and Regulations and elsewhere.)
- Should students in prior grade levels be assessed to identify and address instructional needs and increase the proportion of third, sixth, eighth, and eleventh graders who meet local objectives?
- Does the district's organizational structure suggest grade levels at which it might be particularly useful to assess students—e.g., just before they advance to other school levels?
- Do local curricula or textbooks suggest grade levels at which it might be especially useful to assess students?

Planners may also consider several factors when they decide what time of year to administer each assessment procedure.

- Assessment should be spread out over the school year so that information for improving instruction is available continuously.
- State assessment of third, sixth, eighth, and eleventh grade students occurs in April as specified in the Illinois School Code and elsewhere.
- Districts should allow sufficient time for scoring and data compilation between an assessment procedure's administration and its interpretation and use.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing.* Washington, DC. American Psychological Association. (A copy of this document has been sent to every Illinois school district.)

The most recent edition of this publication includes discussions of validity and reliability which can guide both test selection and test development.

Bloom, B.S. (Ed.) (1956). *Taxonomy of Educational Objectives. The Classification of Educational Goals.* Handbook 1. Cognitive Domain. New York: McKay.

Gronlund, N.E. (1985). *Measurement and Evaluation in Teaching.* New York. MacMillan Publishing Company, Inc.

Validity and reliability are discussed extensively in Part One. Appendix E (Tables E.1 and E.2) includes a description of the major categories in Bloom's taxonomy and illustrations of each.

National Evaluation Systems, Inc. (1987). *Bias Issues in Test Development.* Amherst, MA. National Evaluation Systems, Inc.

This document was prepared under contract with the Illinois State Board of Education. It discusses test bias through language usage, stereotyping, representational unfairness, and content exclusion. It includes guidelines for avoiding bias toward racial, ethnic or religious minority groups, females, and people with disabilities.

Popham, W.J. (1981). *Modern Educational Measurement.* Englewood Cliffs, N.J.. Prentice-Hall, Inc.

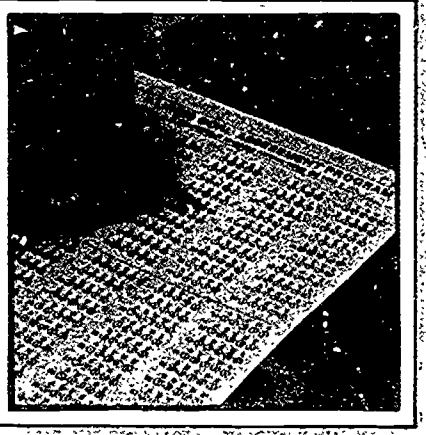
Includes discussion of validity, reliability, test bias, Bloom's taxonomy of cognitive levels, and other assessment topics.

Quellmalz, E.S. (1985, October). Needed. Better Methods for Testing Higher-Order Thinking Skills. *Educational Leadership,* 43(2), 29-35.

Argues that student assessment should require more higher order reasoning and suggests assessment procedures which address that need.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 8

Things to Do Before, During, and After Testing

Often local staff devote much effort to designing assessment programs--selecting standardized tests, developing local tests, and so on. Once programs are designed though, additional thought about how to administer them is needed.

Important details must be attended to before, during, and after tests are administered. Some details can be handled by one person in a district; others require a cooperative effort among administrators, teachers, students, and even parents. Checklists in the chapter appendix outline responsibilities at the district, school, and classroom levels. The information here explains some of those activities further.

Before Testing Occurs

- Determine the number of tests, answer sheets and manuals needed for each grade. Ordering extra tests and answer sheets (say five to ten percent more than needed) for an emergency is a good idea.
- Because of test security, do not stockpile standardized tests. On the other hand, allow plenty of time for receiving materials and for distributing them to schools. Ordering materials three months before planned use should allow enough lead time.
- As materials arrive, check to make sure that everything necessary was received. Open boxes and determine that sufficient quantities of all materials were received.
- As testing time approaches, organize the materials for easy distribution. To preserve security, do not distribute tests to schools until about two weeks before testing, but do arrange them for later distribution and keep them in a locked central location.

*Chapter adapted from Alaska *Assessment Handbook.*

When Testing Occurs

- Be sure that testing rooms have accurate clocks with second hands. If not, test administrators can use stopwatches and write the time on chalkboards at intervals to inform students of how much time remains.
- Read the administration manual and test carefully before the actual testing period starts. Highlight and be prepared to explain any directions that students might not understand.

Practical Tip

Students get cues about how to react to a test from teachers. They need to know the teacher wants the test information. If a teacher reads the directions carefully and proctors the test rigorously, students are more inclined to give their best effort. But if a teacher conducts the testing too casually and haphazardly, paying little attention to the test directions or to class behavior, students may react accordingly and the test will not sufficiently indicate what they know.

After testing is finished:

- If the school or district has a history of lost mail, make copies of the answer sheets before sending them for scoring.
- Clearly specify any special scoring options, summary reports, or data handling requests. Going back later and performing subsequent analyses is much more expensive than doing them the first time through. (Decisions about report contents and table formats should have been made when tests were selected or developed.)

When scores are received:

- Let students know how they did on the test personally, or if only group results are available, how their group performed. Taking a test and then never learning the results can be very frustrating. If appropriate, teachers may use the discussion of test results as an opportunity for review or reteaching.

Preparing Students to Take Tests

The best way to prepare students to take tests is to be sure that they have mastered the content the tests measure. However, research has shown that some students are "testwise"; that is, they have certain skills which are independent of their knowledge of the content tested but which make them better at taking tests. Consequently, there is a small but consistent difference in test scores in favor of students who are testwise.

Testwiseness can be taught, and teaching it is considered ethical. The following list includes suggestions that do not make students any "smarter," but help ensure that they get credit for what they do know.

Testwise students:

- Choose the most correct answer, even if more than one choice is partly true.
- Always estimate the answer for a number problem before working it to see if the final result is reasonable.
- Look at the questions which accompany a reading passage or story problem before reading the passage to notice the information that is needed.
- Watch out for "None of the above" and "All of the above" answer choices. These choices make items harder because students have to decide whether any answers are right or whether more than one is correct.
- Guess if they don't know an answer, especially if they are sure one or more of the responses is obviously wrong.
- Skip the hardest questions on the first pass through a test. Answer the easier items first, then go back and work on the harder ones.
- Carefully mark answers on computer-read answer sheets. Erase marks completely when changing an answer, and do not put stray marks on the sheet. Make marks dark enough for the computer to read.
- Stay aware of the time. Note what time it will be when ten minutes remain; use those ten minutes to review answers.
- Get a good night's sleep before a test, and eat a good breakfast that morning. Avoid drinking a lot of liquids or eating a big meal right before the test.
- Stay calm, relax, and concentrate. A little anxiety about tests is okay, but testwise students do not let tests make them overly tense or upset. (Hill [1980]) discusses the effects of high test anxiety and suggests ways to reduce it.)

What Makes a Test "Standardized"?

A test that is labeled standardized is assumed to be given to all students under the same administration conditions. Those conditions are the ones used when the norm group was tested. If the standard conditions are not met, the comparison of results with the norms is invalid. Several changes in administration procedures could invalidate results, including the following, all of which should be *avoided.*

1. **Reading the directions haphazardly.** The administration manual specifies 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 students respond to 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 may lower students' scores unfairly.

3. **Reading questions aloud that are meant to be read silently by students, defining words in test items, or explaining what items are asking.** Answer procedural questions, but never answer questions about content.

4. **Translating items.** Students who are new to an English-speaking school and usually communicate in another language should probably be excused from taking a test in English because the scores will not accurately represent what they know. Translating items into a student's native language is not acceptable when the student's skills are being compared to a norm group.

In summary, normative information that publishers provide about student 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. Deviation from publishers' instructions will invalidate that information.

Teaching to the Test vs. Teaching the Test

During meetings held before school starts in the fall, a district's teachers pore over copies of their standardized achievement test, which will be administered in April. Some teachers are observed copying items. In another district, teachers spend most of a morning reviewing the objectives which their locally developed criterion-referenced tests (CRTs) measure.

The first instance is clearly unethical. The second is not. In fact, it can be supported. Major differences in these situations explain why the second activity is commendable, while the first is to be guarded against.

In the first situation, standardized *test items* are being studied, while in the second CRT *objectives* are being reviewed. 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.

In reality, teachers in the second situation are actually studying the district's objectives when they review CRT objectives. They are reviewing what should be taught, rather than simply what will be tested.

How Teachers Can Help Students

Just as students can do certain things during tests to show their knowledge to best advantage, teachers can do things to help students get ready for tests. Teachers may prepare for the test in the following ways.

- If the test is untimed, **make certain that students have sufficient time to complete it.** Students should be given as much time as they need—as long as they are making progress.
- If a test is timed, **keep students apprised of how much time is left to complete it.** Writing the time on a chalkboard is better than announcing it.
- **Do the practice items with the group.** Otherwise, many students will ignore them. Practice items are very important, for they show students anything unusual about the wording of test questions and help students mark their answers correctly.
- **Make certain that students take the test seriously.** Without playing on students' anxiety, stress the importance of the test. Present the testing situation with a positive attitude about both the usefulness of the test and students' ability to cope with it.

References

Gronlund, Norman E. (1985). Test Selection, Administration, and Use. Chapter 11 in *Measurement and Evaluation in Teaching,* MacMillan, 320-345.

Includes a section labeled "Administering Published Tests" which presents a concise Test Giver's Checklist. The sections on test selection and test scoring also are useful.

Hill, Kennedy T. (1980). Motivation, Evaluation, and Educational Testing Policy. Chapter 4 in *Achievement Motivation* (Leslie J. Fyans, Jr., Ed.), Plenum, 34-95.

Describes the influence of anxiety and other related factors on student performance on standardized tests. Suggests strategies for helping students deal with the demands of testing.

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.

Discusses test-taking skills from the point of view of the student, including testwiseness and test anxiety. Includes a concise but thorough section on test administration.

Iverson, Grace. (1934, 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 the 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. It inc[illegible]ctions specifically for students and parents, sample tests, and practical hints.

Administering an Assessment Program:*
District-Level Responsibilities

Use this checklist to keep track of district-level responsibilities for administering a successful assessment program. Space is provided at the end of each section for adding other activities.

Before Testing

- ____ Schedule testing dates for district. (Communicate these dates to all interested parties.)
- ____ Determine number of students to be tested.
- ____ Determine number of tests, answer sheets, and test manuals needed (allow 5-10% extra).
- ____ Determine which scoring options, summary reports, or data handling will provide the most useful information.
- ____ 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 the 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.
- ____ Train teachers in interpreting test results to students and parents and in using results for instructional improvement.
- ____ Share test results with concerned groups (parents, school board, newspaper, etc.).
- ____ Review test results to analyze student performance on objectives.
- ____ _____
- ____ _____

*Adapted from Alaska *Assessment Handbook.*

Administering an Assessment Program:*
School-Site Responsibilities

Use this checklist to keep track of school-site responsibilities for administering a successful assessment program. The checklist is divided into two parts. Part I lists activities for a school coordinator. Part II lists activities for each teacher who administers tests. Spaces are provided for adding other activities.

PART I: SCHOOL COORDINATOR

Before Testing

____ Arrange for appropriate testing rooms (adequate seating and light, comfortable temperature and ventilation, no distractions from outside, a clock with a second hand, chalkboard).

____ Distribute test materials to teachers (including stopwatches, if needed).

____ 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.

____ Articulate any problems or concerns to the district coordinator.

____ _____

PART II: TEST ADMINISTRATORS

Before Testing

____ Prepare students by discussing the purpose of the test and teaching test-taking skills.

____ 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.

____ Review the administration manual and test carefully for any directions that may be difficult for students to understand.

____ _____

During Testing

____ Arrange desks to face front of room.

____ 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 the group.

____ If tests are timed, keep students informed of the time left to work.

____ If tests are untimed, allow students as much time as they need to finish their work.

____ Circulate during testing 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 district test coordinator.

____ _____

When Score Reports Are Returned

____ Interpret test results for students and parents.

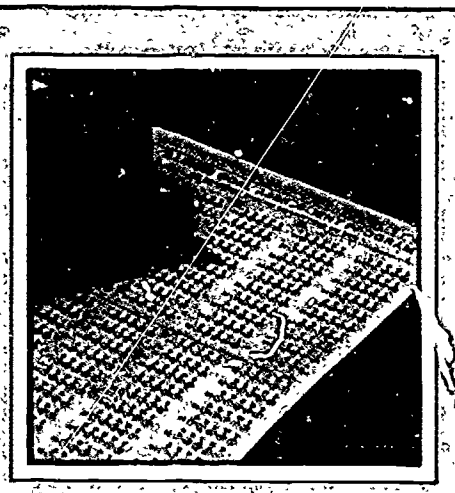
____ Use test results for instructional improvement.

____ _____

*Adapted from Alaska *Assessment Handbook.*

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 9

Why Keep Records?

After assessment scores have been distributed to parents, students, teachers and administrators, what should happen next? What assessment records should be maintained and for how long? What should be kept and what discarded?

Records should not be kept just because "it's always been done." Instead, school personnel should identify reasons for maintaining records and establish recordkeeping procedures appropriate for each information need.

Following are some important reasons for keeping assessment records.

- Parents and students expect schools to keep test scores as part of a student's cumulative record.
- Teachers need information about student achievement to make effective instructional management decisions.
- Districts must comply with the following assurance statement which superintendents sign and submit to the State with the learning assessment plans:

> The district will maintain in a central location a copy of materials related to its Learning Assessment Plan (e.g., learning objectives, descriptions of its assessment procedures, assessment instruments, amendments to this Plan, copies of information disseminated to the public, and correspondence from the State Board of Education approving the Plan and subsequent amendments) to be made available to the State Board of Education for examination upon request.
>
> —Section 210.120(c)(5)

- School and district staff need information about pupil progress to report to the public and to evaluate the effects of curriculum and instruction.

*Chapter adapted from Alaska *Assessment Handbook*.

- 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.

What Types of Information Should Be Kept?

A recordkeeping system needs to be comprehensive enough to maintain 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 district and state tests, as well as student test performance aggregated at the class, school and district levels.

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

- Each school's instructional programs, including curriculum goals and objectives, textbooks, instructional approaches, and any special programs;
- Changes that might be related to student performance, including new curriculum or instructional efforts, new textbooks, changes in teaching staff, changes in instructional time, and shifts in tests or testing procedures.

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

1. Determine all district information needs, at individual student, classroom, school site and district levels;
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 information needs most efficiently and cost-effectively.

How Long Should Assessment Records Be Kept?

One district successfully manages its assessment data using the following guidelines.

- Score reports are returned to students, parents, teachers and administrators as soon 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.
- Grade-level score reports (e.g., group performance summaries) are maintained in the district office until students have graduated. For example, score reports from current 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 score reports from this year's assessment are filed in a single binder).

Using Assessment Records: Three Scenarios

Assessment records, no matter how comprehensive or well maintained, are not very useful by themselves. They become beneficial when data are used to answer important educational questions. The following three examples demonstrate some effective uses of assessment information.

In Evaluation (Scenario 1)

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

District assessment records include information about students' math scores for several years before the new text was adopted, student performance in other subject areas, and student background characteristics. These data will be used with information such as how well the test measures important district goals, how well the test matches the textbook, and teachers' opinions about what has happened. Assessment records will not be the only source of information for this evaluation, but they will make an important contribution.

In Curriculum (Scenario 2)

When assessment scores indicated that study skills was a weak area, a district decided to examine its language arts curriculum. Some definite weaknesses were found in study skills, and new objectives were developed and implemented. The next assessment results showed a definite improvement in study skills although isolated areas remained weak.

Now the curriculum committee will examine the remaining weak areas to determine which warrant more attention. In two years, the newest assessment scores will 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 to use assessment records to help inform educational decisions.

Practical Tip

Many schools have considerable turnover in teaching staff from year to year. New staff should not have to start from scratch to learn their students' current achievement levels. Good assessment records can be extremely valuable to a teacher who wants to know where students stand.

In Instruction (Scenario 3)

Because problem solving is important in a district's mathematics curriculum, a special computer-assisted instructional program is adopted. Equipment is limited; only half the students can enroll. The selection of participants is facilitated by the school's complete records of problem solving scores on the district standardized test. Staff consult these records to identify students who will benefit most from the computerized instruction. Also, teachers are asked to estimate students' problem solving ability Used in combination, teacher judgment and assessment records provide a better instructional decision than either source alone.

Computers Can Help in Recordkeeping

Recordkeeping requires the maintenance and manipulation of data, tasks at which computers excel. Advances in computer technology have made computers affordable to districts that want to purchase computers for handling assessment records. But certain cautions are in order.

- Published, computer-based recordkeeping systems are not always well designed, and even the best of them may not be suitable for a given setting.
- Some computers that are already in schools are not suitable for detailed recordkeeping with many students. Their storage capacities are not sufficient. Before deciding to use a computer system for assessment recordkeeping, districts should make sure that their records will fit.

  Some schools have converted to computer systems, spent months preparing and entering data, and then found that the storage space would not hold all of the records.
- The system may not be flexible enough. Some systems permit only limited types of reports and inquiries.
- 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. The system should do something other than just store information.
- A district may not have the resources to use the system appropriately. For example, if only one set of test scores is not entered, the integrity of the entire system is jeopardized. A computer-based system sometimes makes recordkeeping more formal and intensive than staff are willing or able to undertake.

Large Sites vs. Small

Recordkeeping systems that are practical in districts with large numbers of students may not be practical in smaller districts. An extensive, computer-based system might be very inefficient in a district with only five or ten students per 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). If some information necessary for decision making is not available, it can probably be collected from other existing sources relatively easily.

On the other hand, districts with thousands of students cannot successfully manipulate assessment records by hand, nor can information be collected on an ad hoc basis. The recordkeeping systems for such districts have to be established in advance, probably with the help of experienced data processing personnel. (For information about assessment recordkeeping systems used in the Springfield and Joliet school districts, see documents listed in the reference section.)

Student Records and Compliance with the Law

Administrators need to comply with two federal laws when designing a recordkeeping system. Both the Hatch Amendment, with its associated 1984 Department of Education regulations, and the Family Educational Rights and Privacy Act of 1974 (the Buckley Amendment) affect 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 certain rights 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 an opportunity 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" involving potentially embarrassing psychological problems, antisocial or self-incriminating behavior, criticisms of family members, and statements of family income.

The Buckley Amendment

- Students' educational records must be released upon request to parents (including a noncustodial parent) or to students 18 years of age or older.
- Personally identifiable information in student records may be disclosed only with written approval of parents.
- 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. *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.

Joliet Township High School District #204. (1986). *How to Do It.* (Available at Educational Service Centers.)
Describes the use of a computerized instructional management system which is available commercially.

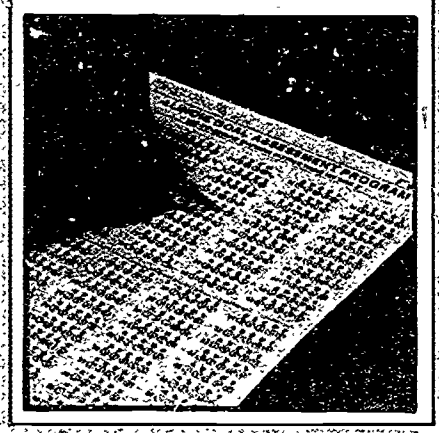
Lyman, Howard B. (1980). Metrics Used in Reporting Test Results. *New Directions in Testing and Measurement; Interpreting Test Performance.*
Discusses in detail the three basic types of scores used to describe test performance. Helpful in deciding how a local district should record its own student scores.

Sampson, James P.; Tenhagen, Carl A., & 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 266-item bibliography.

Springfield Public Schools. (1986). *The Springfield CRT System.* (Available at Educational Service Centers.)
Tells how to use a computerized criterion-referenced testing system.

ASSESSMENT HANDBOOK

A Guide for Assessing Illinois' Students



Chapter 10

Districts are required to develop systems for reporting their assessment results to local residents annually. The reporting systems shall include at least "statements of the degree to which the district's goals, objectives and expectations for student achievement are being met, and if not, what appropriate actions are being taken." That information shall be disseminated through reports presented at or sent to regular school board meetings, local newspapers, and students' parents (from adopted rules for the Learning Assessment and School Improvement Plans, Section 210.130—Reporting System). Districts should make several kinds of decisions about their reporting systems, as discussed this chapter.

Deciding What to Report

Local planning teams should decide what level of detail as well as what kinds of information to report. They should ask questions such as the following.

Will data be summarized across objectives (within goal, learning area, and grade level)?

Will data be reported for each objective separately? Each assessment procedure?

Will data be aggregated across students, classrooms, and schools (i.e., reported at the district level only)?

Will school-level data be reported?

Will the data be aggregated in other ways also?

Will reports include anything other than data on student progress on local expectations? For example, will they include information such as attendance and dropout rates? Norm-referenced test data? State assessment data? Students' responses to sample assessment items?

The data that districts report on student achievement of local objectives may consist of simple percentages of students who met local expectations or

criterion levels. Although norm-referenced tests may be used to assess some objectives, normative scores may not be available for the specific clusters of assessment items used with individual objectives.

Displays of assessment information regarding student achievement of local goals and objectives may be a major feature of a district's report to the public. However, the report might also include other elements such as:

- state assessment data,
- normative data from publishers' standardized shelf tests,
- other indicators of school effectiveness* (e.g., attendance and dropout rates, academic awards, and constituent satisfaction),
- student performance on selected illustrative assessment items/procedures, and
- indicators of local problems (e.g., student mobility, absenteeism, and class size).

Interpreting and Reporting Assessment Data

Because norm-referenced tests have been used widely for many years, considerable information exists about how to interpret and report the results. Some of that information is included in the appendix along with a few guidelines for developing charts.

Interpreting data from other assessment procedures, such as locally constructed tests or performance assessment, and reporting the data to the public may be difficult for local educators. Especially during the first few years, districts may have only rough estimates of the difficulty level of various items and other procedures. Even establishing realistic local expectations may be difficult. Local staff will need to explain these complexities to their public. With experience, they can adjust expectancies and gradually collect data that are increasingly meaningful.

The frequent use of norm referenced tests and the existence of several types of statistically derived scores make it easy to assume that various uses of the scores are warranted. However, staff must exercise caution when using norm referenced test data. The following are several important precautions.

Be extremely careful about aggregating data across tests. Such aggregation is not legitimate (Frechtling & Myerberg, 1983). Although two tests may appear to measure the same learning area, they may do so very differently. The content of the tests may vary widely. The format of the items may be so different that even if the content were the same, student performance might not be consistent across the two tests. Also, the tests were normed on different groups of students. Although publishers attempt to use "nationally representative" samples, the groups sometimes vary considerably.

Do not attempt to aggregate certain types of normative scores. Regardless of whether the comparison is of different groups of students or of a single group at different points in time, small differences may be due more to measurement errors than actual differences in students. All scores contain measurement error due to factors such as ambiguous items, guessing, and distractions during the testing situation. Even "statistically significant" differences (i.e., differences that are supposedly "real" because they are greater than the estimated measurement error) should be interpreted cautiously, especially if they are small or represent very large groups of students. Another reason for interpreting group differences carefully is that they may be due more to external factors than to schools. For example, some groups' cultural backgrounds may have given them more opportunities to acquire certain knowledge/skills outside of school.

Be careful when comparing results across grade levels. Again, the norms are based on different samples which are probably not comparable. Also, differences in test content and fidelity to the curriculum may help explain group differences. As Frechtling and Myerberg (1983) suggest, the test score declines that commonly appear as students progress through school may be due mainly to increased curricular variety and a consequent decline in the alignment between what is taught and what is tested.

References

Frechtling, J.A. & Myerberg, NJ (1983). *Reporting Test Scores to Different Audiences.* (ERIC/TM Report 85.) Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service.

These guidelines for reporting test scores, which also contain discussions of test score interpretation and use, include excerpts and charts from achievement reports prepared by various school districts.

Zion-Benton Township High School District 126. (1986). *School Performance Concepts and Information Plan.* (Available at Educational Service Centers.)

This document includes copies of several types of student achievement reports issued by the district.

*See the Chapter appendix for the measures of program effectiveness used by Zion Benton Township High School.

Program Effectiveness Indicaters Used by Zion-Benton High School

I. ANALYSIS OF EXTERNAL TESTING

—Educational Development Series (EDS) – academic achievement test administered to all students in 8th, 11th, 12th grades

—ACT, SAT, PSAT, AP, NMSQT—voluntary tests taken by college-bound students

II. ANALYSIS OF DEPARTMENTAL TESTING

—Departmental Final Examination Results

—Grade Analysis

III. ANALYSIS OF AWARDS AND INTERSCHOLASTIC COMPETITION PERFORMANCE

—Academic Contests

—Artistic Contests

—Athletic Contests

IV. ANALYSIS OF BEHAVIOR AND DISCIPLINE STATISTICS

—Attendance Patterns

—Discipline Patterns

—Dropout Rate

—Suspensions

—Expulsions

V. ANALYSIS OF CONSTITUENT SATISFACTION

—Parents' Perceptions

—Teachers' Perceptions

—Seniors' Perceptions

—Graduates' Perceptions

Sample Charts for Reporting Test Results

Chart 1 - *Poor*

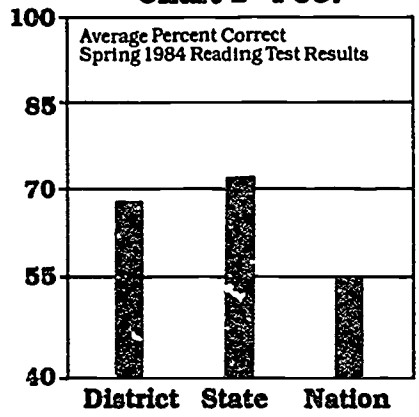
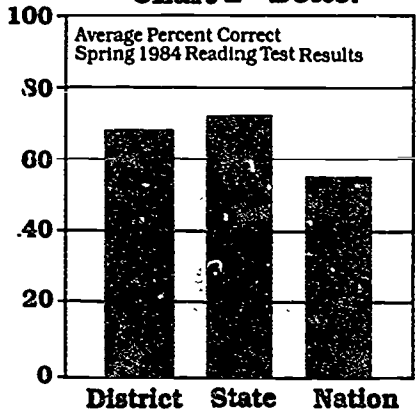


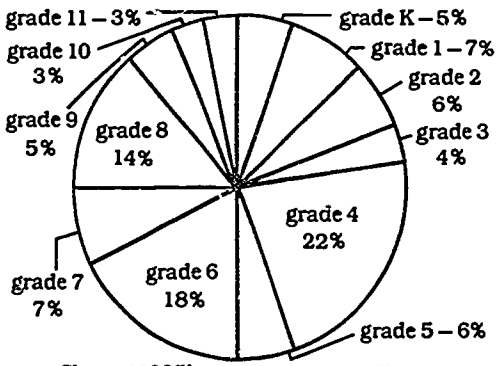
Chart 2 - *Better*



What makes Chart 2 better than Chart 1?

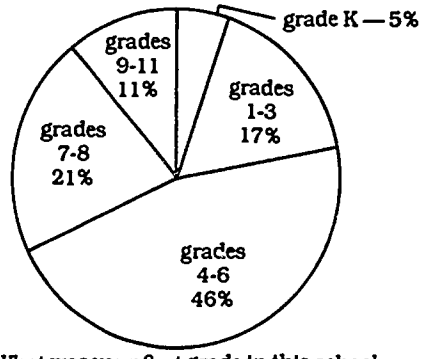
- Bars are wider than the space between them.
- Zero is included on axis. (Zero *can* be omitted, but break in axis should alert reader.)
- Grid lines don't pass through bars.
- Y axis uses scale that makes bars easy to interpret.

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*

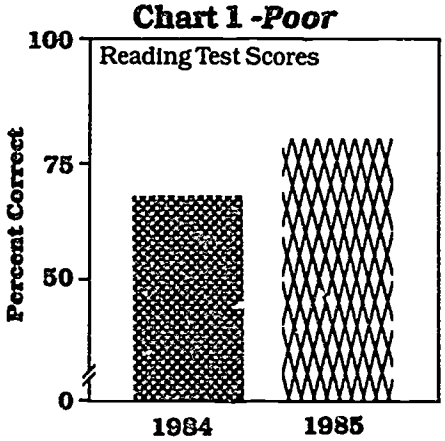
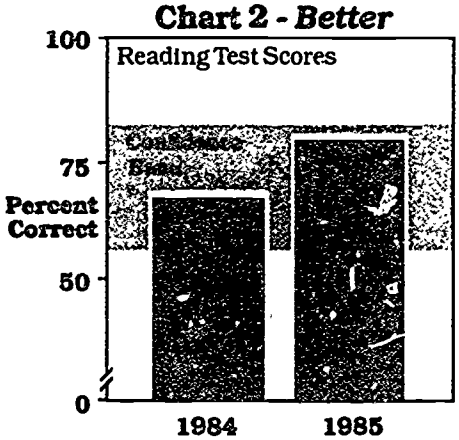


Chart 2 - *Better*



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*

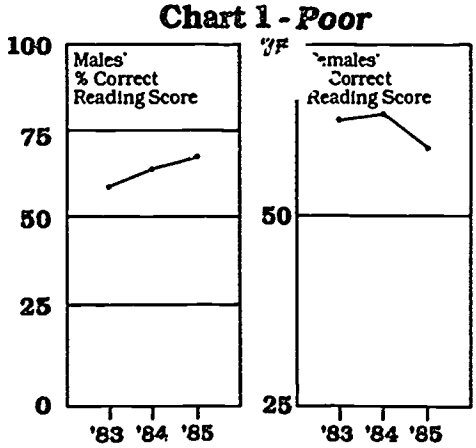
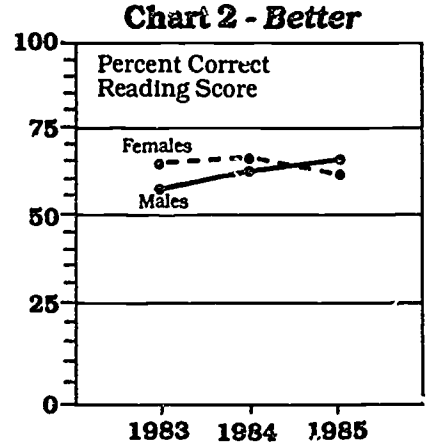


Chart 2 - *Better*



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.

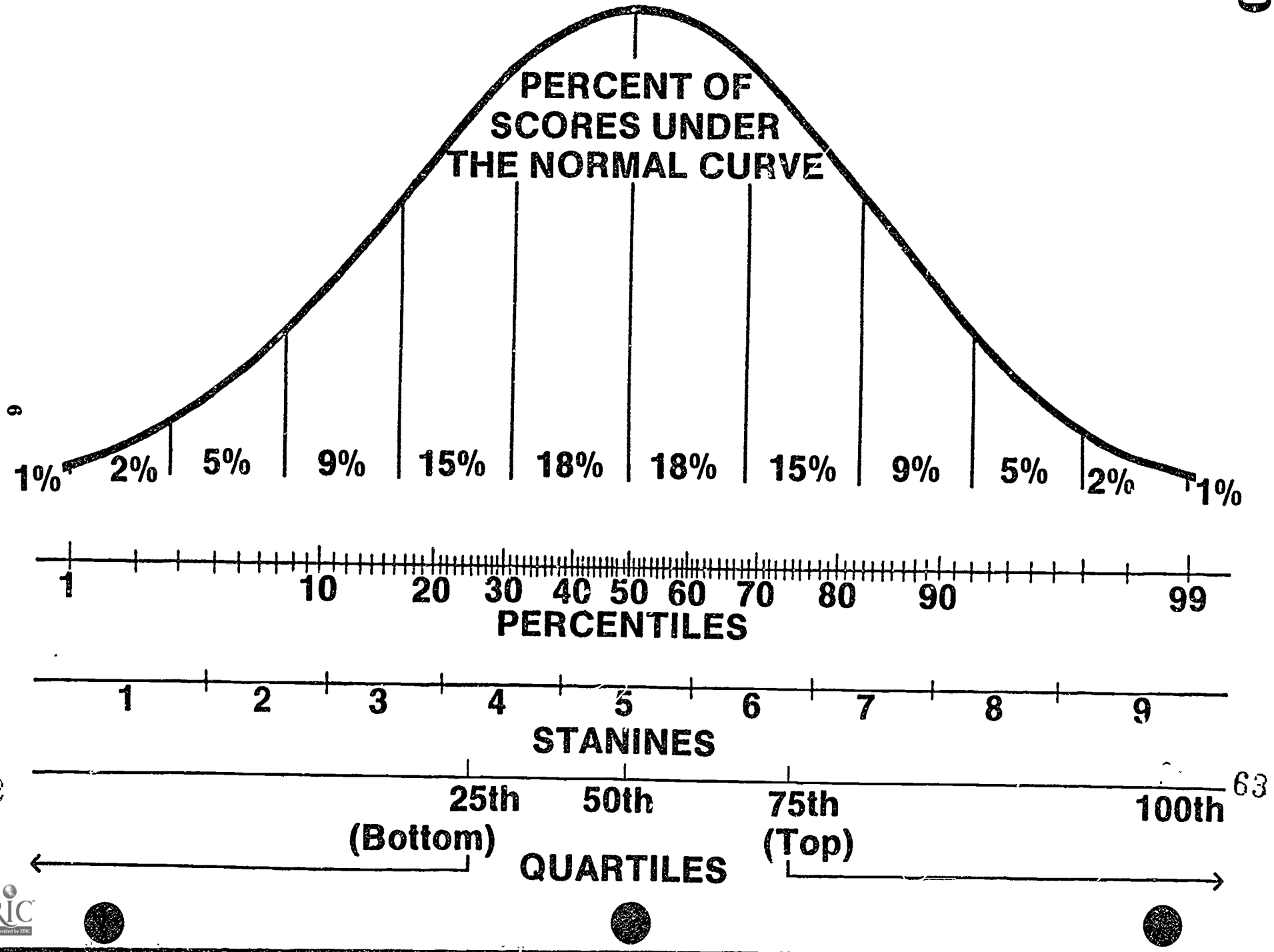
*From Alaska *Assessment Handbook.*

Properties of Different Score Systems

Name of Score	Description	Major Advantages	Major Limitations
Raw Score	Indicates the total number of test items correctly answered by a student.	+Can be computed easily. +Can be used to assess mastery.	-Should not be averaged. -Not suitable for comparing performances on different tests or subtests.
Percentile	Indicates the percentage of students in the norm group whose score is less than or equal to a given score.	+Easy to explain. +Useful for comparing the performance of an individual with a norm group.	-Not an equal interval scale, and so cannot be averaged among students or subject areas.
Stanine	Distributes scores into nine broad intervals.	+Provides a general description of student performance level. +Can be averaged.	-Does not allow fine discrimination.
Normal Curve Equivalent (NCE)	Converts percentiles into a normalized equal interval scale suitable for computing and comparing gains in achievement.	+Has the combined advantages of percentiles and stanines. It can be used for comparing the performance of a group with that of a norm group and can also be meaningfully averaged.	-Is a scale not familiar to many. -Not all test publishers use it.
Expanded Standard Score	Is an equal interval score system which links several overlapping levels of a test.	+Makes it possible to track students longitudinally from grade to grade. +Also possible to test students at functional level and interpret results at their grade level.	-Interpretation requires familarity with particular test and subtest. -Is given different names by different publishers

Prepared by ECIA Chapter I Technical Assistance Center

Appendix D: Chapter 10
Normal Curve Distribution of Test Scores



GENERAL APPENDIX

A Glossary of Assessment Terms*

A through E

- **achievement test:** A test that measures the extent to which a student has acquired certain knowledge or mastered certain skills.

- **alignment:** The process of assuring that curriculum, instruction and assessment procedures all match each other and that communication among educators and administrators at all levels within a district is open and functional. (See Chapter 4)

- **assessment:** The process of estimating, for example, student attainment of learning objectives. A variety of procedures, including tests, can be used in the estimation. (See Chapter 1)

- **comprehensive assessment system:** A planned program for monitoring the achievement of an organization's students. The program is likely to specify at least: a schedule that indicates the grade levels of students who will be assessed, the subject areas, and the time (month) of administration; the assessment procedures that will be used; and the plans for processing, interpreting, and using the results. (See Chapter 1)

- **content validity:** The extent to which a test matches the content of a given program.

- **correlation coefficient:** A measure of the degree of relationship 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.

- **customized test:** A test that is designed specifically for a particular school or district. The test may have been developed by a publisher who maintains a collection of learning objectives and matching assessment items. The test is developed after the district indicates which objectives to assess.

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

- **domain-referenced test (DRT):** Similar to a criterion-referenced test. It is designed to provide information on the extent of student learning in a specific content domain.

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

- **empirical data:** Data collected through observation or experience. The test scores of local students, for example, are empirical data.

- **empirical norm dates:** The actual dates on which a publisher tested the students in the norm group. Publishers recommend that schools administer tests on these dates. Testing at other times may mean that students have received more or less instruction than the norm group.

- **error of measurement:** A statistical estimate of the difference between an observed score and the corresponding "true" score (the score that would be obtained if the assessment were perfectly reliable).

F through N

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

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

- **item banks:** Collections of assessment items. Generally, these are used for constructing tests that measure selected learning objectives. With sufficient numbers of items, multiple test forms that assess the same objectives can be constructed. (See Chapter 6)

- **item-difficulty level:** (See "p value")

- **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 more possible to average scores across groups and to aggregate results across tests.

- **norm group:** The sample of students that was given a test 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

*Adapted from Alaska *Assessment Handbook.*

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 geographic region of the country.

- **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 of an external reference group (the norm group).

- **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 students tested.

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

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

- **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 indicates the percent of students who obtained scores equal to or less than a given score.

- **pilot testing:** Trying out a test or item with a small number of students to see if it works before giving it to a large group of students.

- **point biserial correlation coefficient:** An estimate which is often used to determine whether a test item is functioning well by comparing individual students' performance on the item to their performance on the test as a whole. (Are students with high total scores more likely to answer a question correctly than other students?) Point biserial correlations are between one dichotomous variable (such as whether or not individuals students responded to an item correctly) and one continuous variable (e.g., students' total scores on the test). The range of the coefficients is restricted; generally, correlations above .30 are acceptable.

- **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.

- **reliability:** The extent to which a test can be depended upon to provide consistent, error-free 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. (See Chapter 7)

- **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 from grade to grade or level to level of a test.

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

- **standardized (or uniform) assessment procedure:** A clearly described assessment procedure (for example, a test) with administration directions which were developed so that everyone will administer the procedure in the same way. Student performance will not vary because, for example, administrators give different directions or allow differing lengths of time. (See Chapters 3 and 8)

- **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.

- **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 required for it to be statistically significant. For this reason, many persons talk about both statistical and educational significance when referring to test scores.

- **test specifications:** Descriptions of the distribution of items for a test. These distributions are frequently used during test construction to specify the number or percent of items that assess various content categories. (See Chapter 6)

- **testwiseness:** The possession of skills independent of subject-matter knowledge that make it possible for students to achieve higher test scores. Such skills can be taught and will result in small but consistent improvement in test scores. (See Chapter 8)

- **uniform rating scale:** Standards and criteria for assigning value to, for example, student performance. The purpose of a uniform rating scale is to ensure that values will be consistent across raters. A student's rating would be the same regardless of who did the rating. (See Chapter 3)

- **validity:** The characteristic of a test or other assessment procedure that refers to whether the items in an instrument can provide information that is needed for a particular purpose. Most instruments have multiple validities, one for each likely use of the resulting data. (See Chapter 7)

Appendix B

IMPROVING YOUR TEST QUESTIONS

John C. Ory
Measurement and Evaluation Division
Office of Instructional
and Management Services
University of Illinois
at Urbana-Champaign

Improving Your Test Questions

Choosing the appropriate type of test item to measure students' understanding of course material and their achievement of course goals can often be as difficult a task as writing the items themselves. The purpose of this booklet is (a) to inform you of the uses, advantages and limitations of the various item types and (b) to help you develop specific skills in writing each kind of item.

The booklet is divided into the following sections:

I. Choosing Between Objective and Subjective Test Items

There are two general categories of test items. (a) *objective items* which require students to select the correct response from several alternatives or to supply a word or short phrase to answer a question or complete a statement, and (2) *subjective or essay items* which permit the student to organize and present an original answer. Objective items include multiple choice, true-false, matching and completion, while subjective items include short-answer essay, extended-response essay, problem solving and performance test items. For some instructional purposes one or the other item types may prove more efficient and appropriate. To begin our discussion of the relative merits of each type of test item, test your knowledge of these two item types by answering the following questions.

Test Item Quiz

(circle the correct answer)

1.	Essay exams are easier to construct than are objective exams.	T	F	?
2.	Essay exams require more thorough student preparation and study time than objective exams.	T	F	?
3.	Essay exams require writing skills where objective exams do not.	T	F	?
4.	Essay exams teach a person how to write.	T	F	?
5.	Essay exams are more subjective in nature than are objective exams.	T	F	?
6.	Objective exams encourage guessing more so than essay exams.	T	F	?
7.	Essay exams limit the extent of content covered.	T	F	?
8.	Essay and objective exams can be used to measure the same content or ability.	T	F	?
9.	Essay and objective exams are both good ways to evaluate a student's level of knowledge.	T	F	?

(Quiz answers on following page.)

Quiz Answers

1. TRUE – Essay items are generally easier and less time consuming to construct than are most objective test items. Technically correct and content appropriate multiple choice and true-false test items require an extensive amount of time to write and revise. For example, a professional item writer produces only 9-10 good multiple-choice items in a day's time.

2. ? – According to research findings it is still undetermined whether or not essay tests require or facilitate more thorough (or even different) student study preparation.

3. TRUE – Writing skills do affect a student's ability to communicate the correct "factual" information through an essay response. Consequently, students with good writing skills have an advantage over students who have difficulty expressing themselves through writing.

4. FALSE – Essays do not teach a student how to write but they can emphasize the importance of being able to communicate through writing. Constant use of essay tests may encourage the knowledgeable but poor writing student to improve his/her writing ability in order to improve performance.

5. TRUE – Essays are more subjective in nature due to their susceptibility to scoring influences. Different readers can rate identical responses differently, the same reader can rate the same paper differently over time, the handwriting, neatness or punctuation can unintentionally affect a paper's grade and the lack of anonymity can affect the grading process. While impossible to eliminate, scoring influences or biases can be minimized through procedures discussed later in this booklet.

6. ? – Both item types encourage some form of guessing. Multiple-choice, true-false and matching items can be correctly answered through blind guessing, yet essay items can be responded to satisfactorily through well written bluffing.

7. TRUE – Due to the extent of time required by the student to respond to an essay question, only a few essay questions can be included on a classroom exam. Consequently, a larger number of objective items can be tested in the same amount of time, thus enabling the test to cover more content.

8. TRUE – Both item types can measure similar content or learning objectives. Research has shown that students respond almost identically to essay and objective test items covering the same content. Studies[1] by Sax & Collet (1968) and Paterson (1926) conducted forty-two years apart reached the same conclusion:

> "... there seems to be no escape from the conclusions that the two types of exams are measuring identical things." (Paterson, p. 246)

This conclusion should not be surprising; afterall, a well written essay item requires that the student (1) have a store of knowledge, (2) be able to relate facts and principles, and (3) be able to organize such information into a coherent and logical *written expression*, whereas an objective test item requires that the student (1) have a store of knowledge, (2) be able to relate facts and principles, and (3) be able to organize such information into a coherent and logical choice among several alternatives.

9. TRUE – Both objective and essay test items are good devices for measuring student achievement. However, as seen in the previous quiz answers, there are particular measurement situations where one item type is more appropriate than the other. Following is a set of recommendations for using either objective or essay test items. (Adapted from Robert L. Ebel, *Essentials of Educational Measurement*, 1972, p. 144).

---

[1] Gilbert Sax and LeVerne S. Collet, "An Empirical Comparison of the Effects of Recall and Multiple Choice Tests on Student Achievement," *Journal of Educational Measurement*, vol.5 (1968), 169-73.

Donald G. Paterson, "Do New and Old Type Examinations Measure Different Mental Functions?" *School and Society*, vol. 24 (August, 21, 1926), 246-48.

When to Use Essay or Objective Tests

Essay tests are especially appropriate when:

- the group to be tested is small and the test is not to be reused.
- you wish to encourage and reward the development of student skill in writing.
- you are more interested in exploring the student's attitudes than in measuring his/her achievement.
- you are more confident of your ability as a critical and fair reader than as an imaginative writer of good objective test items.

Objective tests are especially appropriate when:

- the group to be tested is large and the test may be reused.
- highly reliable test scores must be obtained as efficiently as possible.
- impartiality of evaluation, absolute fairness, and freedom from possible test scoring influences (e.g., fatigue, lack of anonymity) are essential.
- you are more confident of your ability to express objective test items clearly than of your ability to judge essay test answers correctly.
- there is more pressure for speedy reporting of scores than for speedy test preparation.

Either essay or objective tests can be used to:

- measure almost any important educational achievement a written test can measure.
- test understanding and ability to apply principles.
- test ability to think critically.
- test ability to solve problems.
- test ability to select relevant facts and principles and to integrate them toward the solution of complex problems.

In addition to the preceding suggestions, it is important to realize that certain item types are *better suited* than others for measuring particular learning objectives. For example, learning objectives requiring the student *to demonstrate* or *to show,* may be better measured by performance test items, whereas objectives requiring the student *to explain* or *to describe* may be better measured by essay test items. The matching of learning objective expectations with certain item types can help you select an appropriate kind of test item for your classroom exam as well as provide a higher degree of test validity (i.e., testing what is supposed to be tested). To further illustrate, several sample learning objectives and appropriate test items are provided on the following page.

Learning Objective	Most Suitable Test Item
The student will be able to categorize and name the parts of the human skeletal system.	Objective Test Item (M-C, T-F, Matching)
The student will be able to critique and appraise another student's English composition on the basis of its organization.	Essay Test Item (Extended-Response)
The student will demonstrate safe laboratory skills.	Performance Test Item
The student will be able to cite four examples of satire that Twain uses in *Huckleberry Finn.*	Essay Test Item (Short-Answer)

After you have decided to use either an objective, essay or both objective and essay exam, the next step is to select the kind(s) of objective or essay item that you wish to include on the exam. To help you make such a choice, the different kinds of objective and essay items are presented in the following section of this booklet. The various kinds of items are briefly described and compared to one another in terms of their advantages and limitations for use. Also presented is a set of general suggestions for the construction of each item variation.

II. Suggestions for Using and Writing Test Items

Multiple-Choice Test Items

The multiple-choice item consists of two parts. (a) the stem, which identifies the question or problem and (b) the response alternatives. Students are asked to select the *one* alternative that best completes the statement or answers the question. For example,

Sample Multiple-Choice Item

(a) *Item Stem: Which of the following is a chemical change?*

(b) *Response Alternatives:*

a. Evaporation of alcohol
b. Freezing of water
**c. Burning of oil*
d. Melting of wax

*correct response

Advantages in Using Multiple-Choice Items

Multiple-choice items can provide

-- versatility in measuring all levels of cognitive ability.
-- highly reliable test scores.
-- scoring efficiency and accuracy.
-- objective measurement of student achievement or ability.
— a wide sampling of content or objectives.
— a reduced guessing factor when compared to true-false items.
— different response alternatives which can provide diagnostic feedback.

Limitations in Using Multiple-Choice Items

Multiple-choice items

— are difficult and time consuming to construct.
— lead an instructor to favor simple recall of facts.
— place a high degree of dependence on the student's reading ability and instructor's writing ability.

Suggestions for Writing Multiple-Choice Test Items

THE STEM

1. When possible, state the stem as a direct question rather than as an incomplete statement.

Undesirable: *Alloys are ordinarily produced by...*

Desirable: *How are alloys ordinarily produced?*

2. Present a definite, explicit and singular question or problem in the stem.

Undesirable: *Psychology...*

Desirable: *The science of mind and behavior is called...*

3. Eliminate excessive verbiage or irrelevant information from the stem.

Undesirable: *While ironing her formal, Jane burned her hand accidentally on the hot iron. This was due to a transfer of heat by...*

Desirable: *Which of the following ways of heat transfer explains why Jane's hand was burned after she touched a hot iron?*

4. Include in the stem any word(s) that might otherwise be repeated in each alternative.

Undesirable: *In national elections in the United States the President is officially*

a chosen by the people.
b. chosen by members of Congress.
c. chosen by the House of Representatives.
**d. chosen by the Electoral College.*

Desirable: *In national elections in the United States the President is officially chosen by*

a. the people.
b. members of Congress.
c. the House of Representatives.
**d. the Electoral College.*

5. Use negatively stated stems sparingly. When used, underline and/or capitalize the negative word.

Undesirable: *Which of the following is not cited as an accomplishment of the Kennedy administration?*

Desirable: *Which of the following is NOT* cited as an accomplishment of the Kennedy administration?

ITEM ALTERNATIVES

6 Make all alternatives plausible and attractive to the less knowledgeable or skillful student.

What process is most nearly the opposite of photosynthesis?

Undesirable	Desirable
a. Digestion	*a. Digestion*
b. Relaxation	*b. Assimilation*
**c. Respiration*	**c. Respiration*
d. Exertion	*d. Catabolism*

7. Make the alternatives grammatically parallel with each other, and consistent with the stem.

Undesirable: *What would do most to advance the application of atomic discoveries to medicine?*

*a. Standardized techniques for treatment of patients.
b. Train the average doctor to apply radioactive treatments.
c. Remove the restriction on the use of radioactive substances.
d. Establishing hospitals staffed by highly trained radioactive therapy specialists.

Desirable: *What would do most to advance the application of atomic discoveries to medicine?*

*a. Development of standardized techniques for treatment of patients.
b. Training of the average doctor in application of radioactive treatments.
c. Removal of restriction on the use of radioactive substances.
d. Addition of trained radioactive therapy specialists to hospital staffs.

8. Make the alternatives mutually exclusive.

Undesirable: *The daily minimum required amount of milk that a 10-year-old child should drink is*

a. 1-2 glasses.
*b. 2-3 glasses.
*c. 3-4 glasses.
d. at least 4 glasses.

Desirable: *What is the daily minimum required amount of milk a 10-year-old child should drink?*

a. 1 glass
b. 2 glasses
*c. 3 glasses
d. 4 glasses

9. When possible, present alternatives in some logical order (e.g., chronological, most to least, alphabetical).

At 7 a.m., two trucks leave a diner and travel north. One truck averages 42 miles per hour and the other truck averages 38 miles per hour. At what time will they be 24 miles apart?

Undesirable	Desirable
a. 6 p.m.	a. 1 a.m.
b. 9 p.m.	b. 6 a.m.
c. 1 a.m.	c. 9 a.m.
*d. 1 p.m.	*d. 1 p.m.
e. 6 a.m.	e. 6 p.m.

10. If you have decided to use a traditional single correct answer format, be sure there is only one correct or best answer.

Undesirable: *The two most desired characteristics in a classroom test are validity and*

a. precision.
*b. reliability.
c. objectivity.
*d. consistency.

Desirable: *The two most desired characteristics in a classroom test are validity and*

a. precision.
**b. reliability.*
c. objectivity.
d. standardization.

11. Make alternatives approximately equal in length.

Undesirable: *The most general cause of low individual incomes in the United States is*

**a. lack of valuable productive services to sell.*
b. unwillingness to work.
c. automation.
d. inflation.

Desirable: *What is the most general cause of low individual incomes in the United States?*

**a. A lack of valuable productive services to sell.*
b. The population's overall unwillingness to work.
c. The nation's increased reliance on automation.
d. An increasing national level of inflation.

12. Avoid irrelevant clues such as grammatical structure, well known verbal associations or connections between stem and answer.

Undesirable: (grammatical clue) *A chain of islands is called an:*

**a. archipelago.*
b. peninsula.
c. continent.
d. isthmus.

Undesirable: (verbal association clue) *The reliability of a test can be estimated by a coefficient of:*

a. measurement.
**b. correlation.*
c. testing.
d. error.

Undesirable: (connection between stem and answer clue) *The height to which a water dam is built depends on*

a. the length of the reservoir behind the dam.
b. the volume of water behind the dam.
**c. the height of water behind the dam.*
d. the strength of the reinforcing wall.

13. Use at least four alternatives for each item to lower the probability of getting the item correct by guessing.

14. Randomly distribute the correct response among the alternative positions throughout the test having approximately the same proportion of alternatives a, b, c, d and e as the correct response.

15. Use the alternatives "none of the above" and "all of the above" sparingly. When used, such alternatives should occasionally be used as the correct response.

True-False Test Items

A true false item can be written in one of three forms. simple, complex, or compound. Answers can consist of only two choices (simple), more than two choices (complex), or two choices plus a conditional completion response (compound). An example of each type of true-false item follows:

Sample True-False Item: Simple

The acquisition of morality is a developmental process. *True* *False*

Sample True-False Item: Complex

The acquisition of morality is a developmental process. *True* *False* *Opinion*

Sample True-False Item: Compound

The acquisition of morality is a developmental process. *True* *False*

If this statement is false, what makes it false?

_____

_____

Advantages in Using True-False Items

True-false items can provide

- the widest sampling of content or objectives per unit of testing time.
- scoring efficiency and accuracy.
- versatility in measuring all levels of cognitive ability.
- highly reliable test scores.
- an objective measurement of student achievement or ability.

Limitations in Using True-False Items

True-false items

- incorporate an extremely high guessing factor. For simple true-false items, each student has a 50/50 chance of correctly answering the item without any knowledge of the item's content.
- can often lead an instructor to write ambiguous statements due to the difficulty of writing statements which are unequivocally true or false.
- do not discriminate between students of varying ability as well as other item types.
- can often include more irrelevant clues than do other item types.
- can often lead an instructor to favor testing of trivial knowledge.

Suggestions for Writing True-False Test Items

1. Base true-false items upon statements that are absolutely true or false, without qualifications or exceptions.

   Undesirable: *Nearsightedness is hereditary in origin.*

   Desirable: *Geneticists and eye specialists believe that the predisposition to nearsightedness is hereditary.*

2. Express the item statement as simply and as clearly as possible.

   Undesirable: *When you see a highway with a marker that reads, "Interstate 80" you know that the construction and upkeep of that road is built and maintained by the state and federal government.*

   Desirable: *The construction and maintenance of interstate highways is provided by both state and federal governments.*

3. Express a single idea in each test item.

   Undesirable: *Water will boil at a higher temperature if the atmospheric pressure on its surface is increased and more heat is applied to the container.*

   Desirable: *Water will boil at a higher temperature if the atmospheric pressure on its surface is increased.*

   and/or

   Water will boil at a higher temperature if more heat is applied to the container.

4. Include enough background information and qualifications so that the ability to respond correctly to the item does not depend on some special, uncommon knowledge.

   Undesirable: *The second principle of education is that the individual gathers knowledge.*

   Desirable: *According to John Dewey, the second principle of education is that the individual gathers knowledge.*

5. Avoid lifting statements from the text, lecture or other materials so that memory alone will not permit a correct answer.

   Undesirable: *For every action there is an opposite and equal reaction.*

   Desirable: *If you were to stand in a canoe and throw a life jacket forward to another canoe, chances are your canoe would jerk backward.*

6. Avoid using negatively stated item statements.

   Undesirable: *The Supreme Court is not composed of nine justices.*

   Desirable: *The Supreme Court is composed of nine justices.*

7. Avoid the use of unfamiliar vocabulary.

Undesirable: *According to some politicians, the raison d'etre for capital punishment is retribution.*

Desirable: *According to some politicians, justification for the existence of capital punishment is retribution.*

8. Avoid the use of specific determiners which would permit a test-wise but unprepared examinee to respond correctly. Specific determiners refer to sweeping terms like "all," "always," "none," "never," "impossible," "inevitable," etc. Statements including such terms are likely to be false. On the other hand, statements using qualifying determiners such as "usually," "sometimes," "often," etc., are likely to be true. When statements do require the use of specific determiners, make sure they appear in both true and false items.

Undesirable: *All sessions of Congress are called by the President. (F)*

The Supreme court is frequently required to rule on the constitutionality of a law. (T)

An objective test is generally easier to score than an essay test. (T)

Desirable: (When specific determiners are used reverse the expected outcomes.)

The sum of the angles of a triangle is always 180°. (T)

Each molecule of a given compound is chemically the same as every other molecule of that compound. (T)

The galvanometer is the instrument usually used for the metering of electrical energy used in a home. (F)

9. False items tend to discriminate more highly than true items. Therefore, use more false items than true items (but no more than 15% additional false items).

Matching Test Items

In general, matching items consist of a column of *stimuli* presented on the left side of the exam page and a column of *responses* placed on the right side of the page. Students are required to match the response associated with a given stimulus. For example,

Sample Matching Test Item

Directions. *On the line to the left of each factual statement, write the letter of the principle which best explains the statement's occurence. Each principle may be used more than once.*

Factual Statements

___ 1. *Fossils of primates first appear in the Cenozoic rock strata, while trilobite remains are found in the Protozoic rocks.*

___ 2. *The Artic and Antarctic regions are sparsely populated.*

___ 3. *Planis have no nervous system.*

___ 4. *Large coal beds exist in Alaska.*

Principles

a. *There have been profound changes in the climate on earth.*

b. *Coordination and integration of action is generally slower in plants than in animals.*

c. *There is an increasing complexity of structure and functions from lower and higher forms of life.*

d. *All life comes from life and produces its own kind of living organism.*

e. *Light is a limiting factor to life.*

Advantages in Using Matching Items

Matching items

- require short periods of reading and response time, allowing you to cover more content.
- provide objective measurement of student achievement or ability.
- provide highly reliable test scores.
- provide scoring efficiency and accuracy.

Limitations in Using Matching Items

Matching items

- have difficulty measuring learning objectives requiring more than simple recall of information.
- are difficult to construct due to the problem of selecting a common set of stimuli and responses.

Suggestions for Writing Matching Test Items

1. Include directions which clearly state the basis for matching the stimuli with the responses. Explain whether or not a response can be used more than once and indicate where to write the answer.

Undesirable: *Directions: Match the following.*

Desirable: *Directions. On the line to the left of each identifying location and characteristics in Column I, write the letter of the country in Column II that is best defined. Each country in Column II may be used more than once.*

2. Use only homogeneous material in matching items.

Undesirable: *Directions: Match the following.*

1. ___ Water	*A. NaCl*
2. ___ Discovered Radium	*B. Fermi*
3. ___ Salt	*C. NH_3*
4. ___ Year of the 1st Nuclear Fission by Man	*D. H_20*
5. ___ Ammonia	*E. 1942*
	F. Curie

Desirable: *Directions. On the line to the left of each compound in Column I, write the letter of the compound's formula presented in Column II. Use each formula only once.*

Column I	*Column II*
1. ___ Water	*A. H_2SO_4*
2. ___ Salt	*B. HCl*
3. ___ Ammonia	*C. NaCl*
4. ___ Sulfuric Acid	*D. H_2O*
	E. H-HCl

3. Arrange the list of responses in some systematic order if possible (e.g., chronological, alphabetical).

Directions: *On the line to the left of each definition in Column I, write the letter of the defense mechanism in Column II that is described. Use each defense mechanism only once.*

Undesirable

Column I	*Column II*
____ 1. *Hunting for reasons to support one's beliefs.*	a. *Rationalization*
____ 2. *Accepting the values and norms of others as one's own even when they are contrary to previously held values.*	b. *Identification*
____ 3. *Attributing to others one's own unacceptable impulses, thoughts and desires.*	c. *Projection*
____ 4. *Ignoring disagreeable situations, topics, sights.*	d. *Introjection*
	e. *Denial of reality*

Desirable

Column II
a. *Denial of reality*
b. *Identification*
c. *Introjection*
d. *Projection*
e. *Rationalization*

4. Avoid grammatical or other clues to the correct response.

Undesirable: *Directions:* *Match the following in order to complete the sentences on the left.*

____ 1. *Igneous rocks are formed*	A. *a hardness of 7.*
____ 2. *The formation of coal requires*	B. *with crystalline rock.*
____ 3. *A geode is filled*	C. *a metamorphic rock.*
____ 4. *Feldspar is classified as*	D. *heat and pressure.*
	E. *through the solidification of molten lava.*

Desirable: Avoid sentence completion due to grammatical clues.

5. Keep matching items brief, limiting the list of stimuli to under 10.

6. Include more responses than stimuli to help prevent answering through the process of elimination.

7. When possible, reduce the amount of reading time by including only short phrases or single words in the response list.

Completion Test Items

The completion item requires the student to answer a question or to finish an incomplete st tement by filling in a blank with the correct word or phrase. For example,

Sample Completion Item

According to Freud, personality is made up of three major systems, the _____, the _____ and the _____.

Advantages in Using Completion Items

Completion items

- can provide a wide sampling of content.
- can efficiently measure lower levels of cognit' ability.
- can minimize guessing as compared to multiple-choice or true-false items.
- can usually provide an objective measure of student achievement or ability.

Limitations in Using Completion Items

Completion items

- are difficult to construct so that the desired response is clearly indicated.
- have difficulty measuring learning objectives requiring more than simple recall of information.
- can often include more irrelevant clues than do other item types.
- are more time consuming to score when compared to multiple-choice or true-false items.
- are more difficult to score since more than one answer may have to be considered correct if the item was not properly prepared.

Suggestions for Writing Completion Test Items

1. Omit only significant words from the statement.

   Undesirable: *Every atom has a central __(core)__ called a nu .us.*

   Desirable: *Every atom has a central core called a(n) __(nucleus)__ .*

2. Do not omit so many words from the statement that the intended meaning is lost.

   Undesirable: *The _____ were to Egypt as the _____ were to Persia and as _____ were to the early tribes of Israel.*

   Desirable: *The Pharaohs were to Egypt as the _____ were to Persia and as _____ were to the early tribes of Israel.*

3. Avoid grammatical or other clues to the correct response.

   Undesirable: *Most of the United States' libraries are organized according to the __(Dewey)__ decimal system.*

   Desirable: *Which organizational system is used by most of the United States libraries? __(Dewey decimal)__*

4. Be sure there is only *one* correct response.

   Undesirable: *Trees which shed their leaves annually are __(seed-bearing, common)__ .*

   Desirable: *Trees which shed their leaves annually are called __(deciduous)__ .*

5. Make the blanks of equal length.

   Undesirable: *In Greek mythology, Vulcan was the son of __(Jupiter)__ and __(Juno)__ .*

   Desirable: *In Greek mythology, Vulcan was the son of __(Jupiter)__ and __(Juno)__ .*

6. When possible, delete words at the end of the statement after the student has been presented a clearly defined problem.

   Undesirable: *__(122.5)__ is the molecular weight of $KClO_3$*

   Desirable: *The molecular weight of $KClO_3$ is __(122.5)__ .*

7. Avoid lifting statements directly from the text, lecture or other sources.

8. Limit the required response to a single word or phrase.

Essay Test Items

The essay test is probably the most popular of all types of teacher made tests. In general, a classroom essay test consists of a small number of questions to which the student is expected to demonstrate his/her ability to (a) recall factual knowledge, (b) organize this knowledge and (c) present the knowledge in a logical integrated answer to the question. An essay test item can be classified as either an extended-response essay item or a short-answer essay item. The latter calls for a more restricted or limited answer in terms of form or scope. An example of each type of essay item follows.

Sample Extended-Response Essay Item

Explain the difference between the S-R (Stimulus-Response) and the S-O-R (Stimulus-Organism-Response) theories of personality. Include in your answer (a) brief descriptions of both theories, (b) supporters of both theories and (c) research methods used to study each of the two theories. (10 pts. 20 minutes)

Sample Short-Answer Essay Item

Identify research methods used to study the S-R (stimulus-response) and S-O-R (Stimulus-Response-Organism) theories of personality. (5 pts. 10 minutes)

Advantages in Using Essay Items

Essay items

- are easier and less time consuming to construct than are most other item types.
- provide a means for testing student ability to compose an answer and present it in a logical manner.
- can efficiently measure higher order cognitive objectives (e.g., analysis, synthesis, evaluation).

Limitations in Using Essay Items

Essay items

- cannot measure a large amount of content or objectives.
- generally provide low test and test scorer reliability.
- require an extensive amount of instructor's time to read and grade.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the grader).

Suggestions for Writing Essay Test Items

1. Prepare essay items that elicit the type of behavior you want to measure.

   Learning Objective: *The student will be able to explain how the normal curve serves as a statistical model.*

   Undesirable: *Describe a normal curve in terms of. symmetry, modality, kurtosis and skewness.*

   Desirable: *Briefly explain how the normal curve serves as a statistical model for estimation and hypothesis testing.*

2. Phrase each item so that the student's task is clearly indicated.

   Undesirable: *Discuss the economic factors which led to the stock market crash of 1929.*

   Desirable: *Identify the three major economic conditions which led to the stock market crash of 1929. Discuss briefly each condition in correct chronological sequence and in one paragraph indicate how the three factors were interrelated.*

3. Indicate for each item a point value or weight and an estimated time limit for answering.

   Undesirable: *Compare the writings of Bret Harte and Mark Twain in terms of settings, depth of characterization, and dialogue styles of their main characters.*

   Desirable: *Compare the writing of Bret Harte and Mark Twain in terms of settings, depth of characterization and dialogue styles of their main characters. (10 points 20 minutes)*

4. Ask questions that will elicit responses on which experts could agree that one answer is better than another.

5. Avoid giving the student a choice among optional items as this greatly reduces the reliability of the test.

6. It is generally recommended for classroom examinations to administer several short-answer items rather than only one or two extended- response items.

Suggestions for Scoring Essay Items

1. Choose a scoring model. Two of the more common scoring models are ANALYTICAL SCORING and GLOBAL QUALITY.

ANALYTICAL SCORING: Each answer is compared to an ideal answer and points are assigned for the inclusion of necessary elements. Grades are based on the number of accumulated points either absolutely (i.e., A = 10 or more points, B = 6-9 pts., etc.) or relatively (A = top 15% scores, B = next 30% of scores, etc.)

GLOBAL QUALITY: Each answer is read and assigned a score (e.g., grade, total points) based either on the total quality of the response or on the total quality of the response relative to other student answers.

Example Essay Item and Grading Models

"Americans are a mixed-up people with no sense of ethical values. Everyone knows that baseball is far less necessary than food and steel, yet they pay ball players a lot more than farmers and steelworkers."

WHY? Use 3-4 sentences to indicate how an economist would explain the above situation.

Analytical Scoring

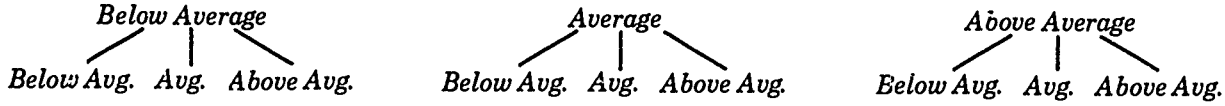
Necessary Elements to be Included in Response	Points
Salaries are based on demand relative to supply of such services.	3
Excellent ball players are rare.	2
Ball clubs have a high demand for excellent players.	2
Clarity of Response.	2
	9 pts.

Global Quality

Assign scores or grades on the overall quality of the written response as compared to an ideal answer. Or, compare the overall quality of a response to other student responses by sorting the papers into three stacks.

Below Average *Average* *Above Average*

Read and sort each stack again and divide into three more stacks.



In total, nine discriminations can be used to assign test grades in this manner. The number of stacks or discriminations can vary to meet your needs.

2. Try *not* to allow factors which are irrelevant to the learning outcomes being measured affect your grading (i.e., handwriting, spelling, neatness).

3. Read and grade all class answers to one item before going on to the next item.

4. Read and grade the answers without looking at the students' names to avoid possible preferential treatment.

5. Occasionally shuffle papers during the reading of answers to help avoid any systematic order effects (i.e., Sally's "B" work always followed Jim's "A" work thus it looked more like "C" work).

6. When possible, ask another instructor to read and grade your students' responses.

Problem Solving Test Items

Another form of a subjective test item is the problem solving or computational exam question. Such items present the student with a problem situation or task and require a demonstration of work procedures and a correct solution, or just a correct solution. This kind of test item is classified as a subjective type of item due to the procedures used to score item responses. Instructors can assign full or partial credit to either correct or incorrect solutions depending on the quality and kind of work procedures presented. An example of a problem solving test item follows.

Example Problem Solving Test Item

It was calculated that 75 men could complete a strip on a new highway in 70 days. When work was scheduled to commence, it was found necessary to send 25 men on another road project. How many days longer will it take to complete the strip? Show your work for full or partial credit.

Advantages in Using Problem Solving Items

Problem solving items

- minimize guessing by requiring the students to provide an original response rather than to select from several alternatives.
- are easier to construct than are multiple-choice or matching items.
- can most appropriately measure learning objectives which focus on the ability to apply skills or knowledge in the solution of problems.
- can measure an extensive amount of content or objectives.

Limitations in Using Problem Solving Items

Problem solving items

- generally provide low test and test scorer reliability.
- require an extensive amount of instructor time to read and grade.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the grader when partial credit is given).

Suggestions for Writing Problem Solving Test Items

1. Clearly identify and explain the problem.

   Undesirable: *During a car crash, the car slows down at the rate of 490 m/sec^2. What is the magnitude and direction of the force acting on a 100-kg driver?*

   Desirable: *During a car crash, the car slows down at the rate of 490 m/sec^2. Using the car as a frame of reference, what is the magnitude and direction of the gram force acting on a 100-kg driver?*

2. Provide directions which clearly inform the student of the type of response called for.

   Undesirable: *An American tourist in Paris finds that he weighs 70 kilograms. When he left the United States he weighed 144 pounds. What was his net change in weight?*

   Desirable: *An American tourist in Paris finds that he weighs 70 kilograms. When he left the United States he weighed 144 pounds. What was his net weight change in pounds?*

3. State in the directions whether or not the student must show his/her work procedures for full or partial credit.

   Undesirable: *A double concave lens is made of glass with n = 1.50. If the radii of curvature of the two lens surfaces are both 30.0 cm, what is the focal length of the lens?*

   Desirable: *A double concave lens is made of glass with n = 1.50. If the radii of curvature of the two lens surfaces are both 30.0 cm, what is the focal length of the lens? Show your work to receive full or partial credit.*

4. Clearly separate item parts and indicate their point values.

   A man leaves his home and drives to a convention at an average rate of 50 miles per hour. Upon arrival, he finds a telegram advising him to return at once. He catches a plane that takes him back at an average rate of 300 miles per hour.

   Undesirable: *If the total traveling time was 1 3/4 hours, how long did it take him to fly back? How far from his home was the convention?*

   Desirable: *If the total traveling time was 1 3/4 hours:*

   (1) How long did it take him to fly back? (1 pt.)

   (2) How far from his home was the convention? (1 pt.)

   Show your work for full or partial credit.

5. Use figures, conditions and situations which create a realistic problem.

   Undesirable: *An automobile weighing 2,840 N (about 640 pounds) is traveling at a speed of 300 miles per hour. What is the car's kinetic energy? Show your work. (2 pts.)*

   Desirable: *An automobile weighing 14,200 N (about 3,200 pounds) is traveling at a speed of 12m/sec. What is the car's kinetic energy? Show your work. (2 pts.)*

6. Ask questions that elicit responses on which experts could agree that one solution and one or more work procedures are better than others.

7. Work through each problem before classroom administration to double-check accuracy.

Performance Test Items

A performance test item is designed to assess the ability of a student to perform correctly in a *simulated situation* (i.e., a situation in which the student will be ultimately expected to apply his/her learning). The concept of *simulation* is central in performance testing, a performance test will simulate to some degree a real life situation to accomplish the assessment. In theory, a performance test could be constructed for any skill and real life situation. In practice, most performance tests have been developed for the assessment of vocational, managerial, administrative, leadership, communication, interpersonal and physical education skills in various simulated situations. An illustrative example of a performance test item is provided below.

Sample Performance Test Item

Assume that some of the instructional objectives of an *urban planning* course include the development of the student's ability to effectively use the principles covered in the course in various "real life" situations common for an urban planning professional. A performance test item could measure this development by presenting the student with a specific situation which represents a "real life" situation. For example,

> *An urban planning board makes a last minute request for the professional to act as consultant and critique a written proposal which is to be considered in a board meeting that very evening. The professional arrives before the meeting and has one hour to analyze the written proposal and prepare his critique. The critique presentation is then made verbally during the board meeting, reactions of members of the board or the audience include requests for explanation of specific points or informed attacks on the positions taken by the professional.*
>
> *The performance test designed to simulate this situation would require that the student to be tested role play the professional's part, while students or faculty act the other roles in the situation. Various aspects of the "professional's" performance would then be observed and rated by several judges with the necessary background. The ratings could then be used both to provide the student with a diagnosis of his/her strengths and weaknesses and to contribute to an overall summary evaluation of the student's abilities.*

Advantages in Using Performance Test Items

Performance test items

- can most appropriately measure learning objectives which focus on the ability of the students to apply skills or knowledge in real life situations.
- usually provide a degree of test validity not possible with standard paper and pencil test items.
- are useful for measuring learning objectives in the psychomotor domain.

Limitations in Using Performance Test Items

Performance test items

- are difficult and time consuming to construct.
- are primarily used for testing students individually and not for testing groups. Consequently, they are relatively costly, time consuming, and inconvenient forms of testing.

- generally provide low test and test scorer reliability.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the observer/grader).

Suggestions for Writing Performance Test Items

1. Prepare items that elicit the type of behavior you want to measure.
2. Clearly identify and explain the simulated situation to the student.
3. Make the simulated situation as "life-like" as possible.
4. Provide directions which clearly inform the students of the type of response called for.
5. When appropriate, clearly state time and activity limitations in the directions.
6. Adequately train the observer(s)/scorer(s) to ensure that they are fair in scoring the appropriate behaviors

III. Checklist for Evaluating Test Items

EVALUATE YOUR TEST ITEMS BY CHECKING THE SUGGESTIONS WHICH YOU FEEL YOU HAVE FOLLOWED.

Multiple-Choice Test Items

_____ When possible, stated the stem as a direct question rather than as an incomplete statement.

_____ Presented a definite, explicit and singular question or problem in the stem.

_____ Eliminated excessive verbiage or irrelevant information from the stem.

_____ Included in the stem any word(s) that might have otherwise been repeated in each alternative.

_____ Used negatively stated stems sparingly. When used, underlined and/or capitalized the negative word(s).

_____ Made all alternatives plausible and attractive to the less knowledgeable or skillful student.

_____ Made the alternatives grammatically parallel with each other, and consistent with the stem.

_____ Made the alternatives mutually exclusive.

_____ When possible, presented alternatives in some logical order (e.g., chronologically, most to least).

_____ Made sure there was only one correct or best response per item.

_____ Made alternatives approximately equal in length.

_____ Avoided irrelevant clues such as grammatical structure, well known verbal associations or connections between stem and answer.

_____ Used at least four alternatives for each item.

_____ Randomly distributed the correct response among the alternative positions throughout the test having approximately the same proportion of alternatives a, b, c, d, and e as the correct response.

_____ Used the alternatives "none of the above" and "all of the above" sparingly. When used, such alternatives were occasionally the correct response.

True-False Test Items

_____ Based true-false items upon statements that are absolutely true or false, without qualifications or exceptions.

_____ Expressed the item statement as simply and as clearly as possible.

_____ Expressed a single idea in each test item.

_____ Included enough background information and qualifications so that the ability to respond correctly did not depend on some special, uncommon knowledge.

_____ Avoided lifting statements from the text, lecture or other materials.

_____ Avoided using negatively stated item statements.

_____ Avoided the use of unfamiliar language.

_____ Avoided the use of specific determiners such as "all," "always," "none," "never," etc., and qualifying determiners such as "usually," "sometimes," "often," etc.

_____ Used more false items than true items (but not more than 15% additional false items).

Matching Test Items

_____ Included directions which clearly stated the basis for matching the stimuli with the response.

_____ Explained whether or not a response could be used more than once and indicated where to write the answer.

_____ Used only homogeneous material.

_____ When possible, arranged the list of responses in some systematic order (e.g., chronologically, alphabetically).

_____ Avoided grammatical or other clues to the correct response.

_____ Kept items brief (limited the list of stimuli to under 10).

_____ Included more responses than stimuli.

_____ When possible, reduced the amount of reading time by including only short phrases or single words in the response list.

Completion Test Items

_____ Omitted only significant words from the statement.

_____ Did not omit so many words from the statement that the intended meaning was lost.

_____ Avoided grammatical or other clues to the correct response.

_____ Included only one correct response per item.

_____ Made the blanks of equal length.

_____ When possible, deleted the words at the end of the statement after the student was presented with a clearly defined problem.

_____ Avoided lifting statements directly from the text, lecture or other sources.

_____ Limited the required response to a single word or phrase.

Essay Test Items

_____ Prepared items that elicited the type of behavior you wanted to measure.

_____ Phrased each item so that the student's task was clearly indicated.

_____ Indicated for each item a point value or weight and an estimated time limit for answering.

_____ Asked questions that elicited responses on which experts could agree that one answer is better than others.

_____ Avoided giving the student a choice among optional items.

_____ Administered several short-answer items rather than 1 or 2 extended-response items.

Grading Essay Test Items

_____ Selected an appropriate grading model.

_____ Tried not to allow factors which were irrelevant to the learning outcomes being measured to affect your grading (e.g., handwriting, spelling, neatness).

_____ Read and graded all class answers to one item before going on to the next item.

_____ Read and graded the answers without looking at the student's name to avoid possible preferential treatment.

_____ Occasionally shuffled papers during the reading of answers.

_____ When possible, asked another instructor to read and grade your students' responses.

Problem Solving Test Items

_____ Clearly identified and explained the problem to the student.

_____ Provided directions which clearly informed the student of the type of response called for.

_____ Stated in the directions whether or not the student must show work procedures for full or partial credit.

_____ Clearly separated item parts and indicated their point values.

_____ Used figures, conditions and situations which created a realistic problem.

_____ Asked questions that elicited responses on which experts could agree that one solution and one or more work procedures are better than others.

_____ Worked through each problem before classroom administration.

Performance Test Items

_____ Prepared items that elicit the type of behavior you wanted to measure.

_____ Clearly identified and explained the simulated situation to the student.

_____ Made the simulated situation as "life-like" as possible.

_____ Provided directions which clearly inform the students of the type of response called for.

_____ When appropriate, clearly stated time and activity limitations in the directions.

_____ Adequately trained the observer(s)/scorer(s) to ensure that they were fair in scoring the appropriate behaviors.

IV. References for Further Reading

Ebel, Robert L. 1965. *Measuring Educational Achievement.* Englewood Cliffs, New Jersey. Prentice Hall, Chapters 4-6.

Ebel, Robert L. 1972. *Essentials of Educational Measurement.* Englewood Cliffs, New Jersey. Prentice Hall, Chapters 5-8.

Gronlund, N.E. 1976. *Measurement and Evaluation in Teaching.* New York. Macmillan Publishing Co., Chapters 6-9.

Mehrens, W.A., and Lehmann, I.J. 1973. *Measurement and Evaluation in Education and Psychology.* New York: Holt, Rinehart & Winston, Inc. Chapters 7-10.

Nelson, C.H. 1970. *Measurement and Evaluation in the Classroom.* New York. Macmillan Publishing Co., Chapters 5-8.

Payne, David A. 1974. *The Assessment of Learning.* Lexington, Mass.. D.C. Heath and Co., Chapters 4-7.

Scannell, D.P. and Tracy, D.B. 1975. *Testing and Measurement in the Classroom.* New York. Houghton Mifflin Co., Chapters 4-6.

Thorndike, R.L. (Ed.) *Educational Measurement* (2nd ed.). Washington, D.C.. American Council on Education, 1971, Chapter 9 (Performance Testing) and Chapter 10 (Essay Exams).

Appendix C

ILLINOIS STATE BOARD OF EDUCATION
DEPARTMENT OF SCHOOL IMPROVEMENT SERVICES
Learning Assessment and School Improvement Plans
Funded Projects

The State Board of Education issued requests for proposals to develop and disseminate effective practices in the Learning Assessment and School Improvement Plan process. Contact persons and descriptions of the pro jects funded for this purpose are included in this list. Persons who want additional information may contact the projects directly.

1.

District: Belleville Public School District 118*
Contact: Ronald B. Riegle, Assistant Superintendent
Address: 105 West "A" Street, Belleville, Illinois 62220
Phone: 618/233-2830
Title: Instructional Monitoring System
Abstract: The assessment system utilizes locally developed exit tests for grades 3, 6, and 8 corresponding to district-developed skills continuums which represent skills and concepts within learning areas and across grade levels. Test results are summarized in computer generated printouts for individual students, classes, attendance centers, and the district. Procedures for use of the data are also built into the system.

2.

District: Brookwood School District 167
Contact: Margot Schlenker, Principal
Address: 201 Glenwood-Dyer Road, Glenwood, Illinois 60425
Phone: 312/758-5190
Title: School Improvement Plan in Language Arts
Abstract: This project for evaluating student writing involves a format and process for analyzing, compiling and reporting results of state reading and writing assessments and using item analysis to identify school and student strengths and weaknesses.

3.

District: Chicago Public Schools
Contact: Dr. Josue M. Gonzalez, Director, Bureau of Resource Development
Address: 1819 West Pershing Road, Chicago, Illinois 60609
Phone: 312/890-8020
Title: A Proposal to Develop the Achievement Component for Assessment of Student Performance in the Fine Arts
Abstract: This project is designed to provide instructionally relevant information to teachers and administrators regarding a child's artistic aptitude. An instrument for assessing a child's progress on (visual) art objectives is to be developed with these characteristics. teacher administered and scored, objective and curricular based, unbiased and fair, reliable and valid, inexpensive and convenient, theoretically sound and rational.

*Written reports about these projects are available at the 18 Educational Service Centers. Information on the other projects will become available at a later date.

4.

University: College of Lake County
Contact: David Ross, Counselor
Address: 19351 West Washington Street, Grayslake, Illinois 60030
Phone: 312/223-6601, Ext. 352
Title: A Regional Assessment and Reporting System
Abstract: This project develops and implements a regional assessment and reporting system for grades 8 through 14, with particular attention on grades 8, 10, and 12 and post high school follow up. Information is compiled by school and grade level from student assessments that present a profile of academic skills attainment, career interests and educational plans. Various educational options are outlined for each student.

5.

District: Evanston Township High School District 202
Contact: Dr. James E. Phillips, Assistant Superintendent
Address: 1600 Dodge Avenue, Evanston, Illinois 60204
Phone: 312/492-3800
Title: Alignment of Local Objectives with State Goals, Individual Student Improvement Plans, Coordination of School Reforms at the Local Level
Abstract: This project addresses the alignment of local objectives with State Goals using a matrix as a visual framework, prescriptive activities for students whose performance is one grade or more below current placement, scheduling techniques and strategies commonly used in business to organize comprehensive school improvement.

6.

District: Glen Ellyn Community Consolidated School District 89
Contact: Nan H. Sevy, Assistant Superintendent for Instruction
Address: 789 Sheehan Avenue, Glen Ellyn, Illinois 60137
Phone: 312/469-8900
Title: Development and Dissemination of Learning Assessment and School Improvement Plan—Effective Practices
Abstract: This system builds a comprehensive calibrated (Rasch-scaled) bank of items designed to test local learning objectives with criterion-referenced tests. A microcomputer program is used to apply the Rasch model to analyze and calibrate test items and a data base management system to store and retrieve items.

7.

District: Grundy Kendall Educational Service Region
Contact: Rob Molek, Special Projects Consultant
Address: 105 Sage Street, Channahon, Illinois 60410
Phone: 815/467-4048
Title: Alignment of Local Objectives with State Goals, Item Analysis, Comprehensive Assessment System
Abstract: This project provides a uniform, comprehensive training program on learning objectives/assessment which contains various segments in writing objectives and developing curriculum, managing curriculum and assessment by using computer software, reviewing and determining computer assessment systems, collecting and disseminating test items for item banks.

8.

District: Hamilton County Community Unit School District 10
Contact: J. Kenneth Hill, Superintendent
Address: 109 North Washington, McLeansboro, Illinois 62859
Phone: 618/643-2328
Title: Computer-Assisted Profile Data Collection and Analysis for Elementary Schools
Abstract: The project develops templates for a school district profile using AppleWorks, a software package which provides a database appropriate for small elementary attendance centers. Three profile areas are included: student demographic data, student grade data, and standardized achievement test data.

9.

District: Herrin Community Unit District 4*
Contact: Harry J. Revelle, Assistant Superintendent
Address: 700 North 10th Street, Herrin, Illinois 62948
Phone: 618/988-8024
Title: Student Mastery Through Instructional Improvement and Individual Planning
Abstract: This assessment system promotes student mastery of learning objectives through instructional improvement and individual educational planning. Special emphasis is on identifying "at risk" students and helping them master the basic skills.

10.

District: Homewood-Flossmoor High School District 233
Contact: Mrs. Leslie R. Wilson, Project Director
Address: 999 South Kedzie Avenue, Flossmoor, Illinois 60422-2299
Phone: 312/799-3000, Ext. 151
Title: A Prototype for Determining Validity and Reliability on Locally Developed Assessments
Abstract: The goal of this project is to identify a process of establishing validity and reliability for locally designed tests of performance and knowledge in any K-12 learning area. A step-by-step procedure determines reliability and validity of various types of locally developed assessments. Language arts, fine arts, and physical development skills are the sample domain.

11.

University: Illinois State University
Contact: Maurice Scharton, Associate Professor
Address: 104 Lawrence, Normal, Illinois 61761
Phone: 309/438-7100
Title: Writing Assessment
Abstract: The purpose of this project is to develop writing assessment materials and workshops for local districts in adapting state writing assessment methods to the measurement of local objectives. The workshops provide administrators and teachers with experience and ability to design assessment prompts (assignments), administer an assessment, develop a scoring guide from sample essays, and score the samples both analytically and holistically.

12.

District: Joliet Township High School District 204*
Contact: Harold Miller, Assistant Superintendent for Instructional Services
Address: 201 East Jefferson Street, Joliet, Illinois 60432
Phone: 815/727-6981
Title: Development of a Common Testing System for Dual Districts
Abstract: The purpose of this project is to determine testing needs and involve teachers, using a "town meeting" model, in the selection and development of a broad-based program for learning assessment and school improvement in the high school district and its nine feeder districts.

13.

District: Kildeer Countryside C.C.S.D. 96
Contact: Thomas W. Many, Assistant Superintendent for Instruction
Address: 777 Checker Drive, Buffalo Grove, Illinois 60089
Phone: 312/459-4260
Title: Developing a Comprehensive Assessment Program for Dual District School System
Abstract: This project involves development of assessment systems in dual district settings with special emphasis on item banks linked directly to objectives. The process assists school personnel in identifying key instructional goals, developing appropriate learning objectives, writing assessment items, designing valid and reliable item banks, and reviewing student achievement through consistent item-analysis procedures.

14.

District: Macon County Regional Office of Education
Contact: Neal Loveall, Assistant Regional Superintendent
Address: 2240 East Geddes Avenue, Decatur, Illinois 62526
Phone: 217/424-3403
Title: Helping Education Through Fine Arts Assessment
Abstract: Coded observation tools for use in assessing K 6 student learning in visual art and creative dance/movement are developed in this project. The observation tools are field-tested, and non-arts teachers are trained in their use.

15.

District: Mt. Prospect Township High School District 214
Contact: Marilynn J. Kulieke, Director of Research and Testing
Address: 799 West Kensington Road, Mt. Prospect, Illinois 60056
Phone: 312/259-5300
Title: Development of an Assessment Banking and Item Analysis System
Abstract: This project develops a computerized assessment banking system which incorporates the elements lacking in commercial software. The assessment banking system, although generic to all learning areas, is operational in mathematics.

16.

District: Naperville Community School District 203
Contact: Michael J. Harkins, Director of Program Development, Evaluation and Research
Address: 655 South Webster Street, Naperville, Illinois 60566
Phone: 312/420-6558
Title: Comprehensive Expressive Arts Assessment
Abstract: This project is designed to research, develop, test and disseminate a nontraditional Fine Arts assessment. Several different types of nontraditional assessment procedures are developed to match Fine Arts goals and objectives.

17.

District: Salem Community High School District 600
Contact: Dale Guthrie, Media Director
Address: Route 37 North, Salem, Illinois 62881
Phone: 618/548-0727
Title: Cooperative for Effective School Reforms
Abstract: A K-12, outcome-based education program involves content reading strategies which promote critical thinking skills. The proposed process combines school improvement activities such as outcome-based education, critical-thinking strategies and improvement of school report card results.

18.

District: Schaumburg Community Consolidated District 54
Contact: Dr. William Kritzmire, Superintendent
Address: 524 East Schaumburg Road, Schaumburg, Illinois 60194
Phone: 312/885-6700
Title: Data-Based School Improvement Process
Abstract: This eight-step, data-based change process is carefully structured for analyzing and applying data to make more informed decisions. Data gathered and measured against the district's criteria provide the basis for the team to target its improvement effort, plan priorities, and identify and implement a step-by-step plan.

19.

District: Southwestern District 9
Contact: Donald J. Stuckey, Associate Superintendent
Address: Post Office, Piasa, Illinois 62079
Phone: 618/729-3221
Title: An Effective Practices in Mathematics and Biological Sciences Learning Assessment and School Improvement
Abstract: This project integrates the K 6 learning objectives in mathematics and sciences and builds an assessment based on those integrated objectives. The process is implemented through an in structional materials approach.

20.

District: Sparta Community Unit District 140*
Contact: Jim Macri, Administrative Assistant
Address: 123 West College Street, Sparta, Illinois 62286
Phone: 618/443-3622
Title: Outcome-Based Educational Program
Abstract: The district has implemented an outcome based educational program (OBE) which relies heavily on the mastery instructional strategies of Bloom and Hunter. The assessment system includes a communication network among staff, pilot teaching of units utilizing mastery strategies, documentation of student achievement on pilot units, and monitoring the impact of OBE implementation.

21.

District: Springfield School District 186*
Contact: Dr. Robert C. Hill, Director of Instruction
Address: 1900 West Monroe, Springfield, Illinois 62704
Phone: 217/525-3026
Title: Comprehensive Assessment System, Nondiscriminatory Locally Designed Test Instruments, Item Analysis; Scoring of Instruments; School Improvement
Abstract: This project addresses the development of a comprehensive assessment system drawing on the district's language arts assessment program; an analysis of locally developed tests for possible discriminatory characteristics and correcting the tests, procedures for analyzing items on locally developed criterion referenced tests using a computerized criterion-referenced test scoring system, scoring criterion-referenced tests using traditional and non-traditional methods, use of assessment results to bring about district based or school based school improvement.

22.

District: Urbana School District 116*
Contact: Don Holste, Associate Superintendent
Address: 1602 South Anderson Street, P. O. Box 3039, Urbana, Illinois 61801
Phone: 217/384-3636
Title: District Assessment Plan
Abstract: Features of the system include a plan for curriculum development, staff development efforts related to that plan, assessment reflecting learner outcomes, multi-year student profiles, a systematic instructional effort to reduce student test anxiety and a clinical reading/writing lab program to assess tenth grade students.

23.

District: Western Community Unit School District 306
Contact: Larry Marsh, Superintendent
Address: Post Office Box 248, Sheffield, Illinois 61361
Phone: 815/454-2503
Title: A Cooperative Approach which Integrates the Scoring of Assessment Instruments Providing Exemplary Reporting Procedures and Validity and Reliability of Locally Designed Testing Instruments
Abstract: The project generates a generic set of materials for use by similar districts in their LAP/SIP process. Trained consultants deliver the process within the Learning Assessment Cooperative and beyond as requested. Phase I focuses on technology applications to the development and alignment of assessment instruments, phase II on reporting, materials development, and consultants.

24.

District: Woodstock Community Unit District 200*
Contact: Thomas W. Reimer, Assistant Principal
Address: 501 West South Street, Woodstock, Illinois 60098
Phone: 815/338-8200
Title: Student Assessment System
Abstract: The assessment system is a cost effective method for identifying concretely stated expectations and determining if those expectations are being met. Expectancies were developed and defined as outcomes to be mastered by the average student by the completion of a grade level. Teachers used the lists of expectancies in developing assessment instruments.

25.

District: Zion Elementary District 6*
Contact: Eugene H. Latz, Superintendent
Address: 1716 27th Street, Zion, Illinois 60099
Phone: 312/872-5455
Title: District Assessment System
Abstract: The district assessment system demonstrates congruence between the written, taught, and tested curriculum; it identifies the goals for learning but also specifies the emphasis given to those goals. The system is an instructional tool to evaluate progress and provide diagnostic information, identify instructional resources for remediation, evaluate instructional programs and provide information for curricular planning, and provide accountability to the community, school board and state.

26.

District: Zion-Benton Township High School District 126*
Contact: David H. Cox, Superintendent
Address: 1606 West 23rd Street, Zion, Illinois 60099
Phone: 312/746-1202
Title: The School Performance Concepts and Information Plan
Abstract: To go beyond student test data in monitoring the educational processes, the district collects and analyzes five categories of information: external testing (e.g., achievement tests, SAT and ACT results, and other test results); local departmental assessments (final examination results and a grade distribution analysis), awards and interscholastic competition performance; behavior and discipline (attendance, expulsions, suspensions and dropouts), and constituent satisfaction (opinions of parents, teachers, students and graduates on the success of the district).

Projects Referenced by Topic

Illinois State Board of Education

100 North First Street
Springfield, Illinois 62777

Thomas Lay Burroughs, Chairman
Illinois State Board of Education

Ted Sanders
State Superintendent of Education

EXCELLENT SCHOOLS: THE ILLINOIS ADVANTAGE