Issues in Achievement Testing

Eva L Baker

CSE Resource Paper No. 3
1982

Center for The Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

TABLE OF CONTENTS

INTRODUCTION

This booklet is intended to help school personnel, parents, students and members of the community understand concepts and research relating to testing in public schools. What drier topic than tests could there be? But what issue has captured the attention of the public and, at the same time, inspired so much anxiety? Even the least interested citizen must be aware of the growing use of testing in American public education. And tests are used for several purposes. At many schools today, children are expected to pass numerous formal tests. Their performance is believed to indicate students' level of academic achievement, and in turn, to reflect the quality of teaching and organization of the school.

Tests are important because they fulfill three general functions. First, they allow for some aspects of education to become public. Test findings are often reported in the newspapers and frequently find their way into political discussion on local, state, and national levels. Even though we will later address the degree to which findings do provide public access to the products of education, they nonetheless are assumed to permit insight into the quality of educational efforts. This insight relates closely to accountability, which is the second general use of tests. Test performance allows the public not only to monitor the schools but to act collectively to assign responsibility and to express expections for improvement. In addition, as schooling has come to serve many diverse purposes, testing has come to be seen as a mechanism for pulling out from the complicated curriculum those areas regarded to be of most significance. In this way, people have assumed that having tests assures that the schools have standards of quality, and this implicit setting of standards is the third general reason for testing. These three general uses--public access, accountability, and standards--relate to testing in its broadest conception.

Tests are also assumed to be useful for specific purposes. In this discussion of tests, only achievement tests will be treated. Such tests are used presumably to assess the subject matter and skills that students have learned. Other types of tests--for example, aptitude tests (used to predict what areas a student might be good in) or attitudinal measures (used to find out how a student feels about self, school, and others)--are popular but do not have the extended use of achievement tests. (For the reader interested in following up on a range of testing issues, there is a selected bibliography and glossary at the end of this booklet.)

The specific functions of tests are many and may be divided into two categories: (1) tests whose results directly affect the students who take them; and (2) tests whose findings affect the instructional program.

TEST USE WITH DIRECT EFFECTS ON STUDENTS

There are at least three kinds of tests in the category of direct effects:

- Tests of certification
- Tests of selection
- Tests for placement

Tests of Certification

Certification tests allow students to become credentialed. Certification tests have long functioned in the professional fields, such as law, where a Bar Examination must be passed, or in medicine, where state certification examinations are required. In many places, teachers are required to pass a test to be certified. Certification has always operated in an informal way in the schools. Passing a final examination in algebra meant that a student was acknowledged as having necessary skills and concepts taught in that course. What is new in education, at least in the intensity and speed with which they are being implemented, are _formal_ tests for certification of public education. Called "competency" or "proficiency" tests, these measures are administered to certify minimum skills acquired by students. Sometimes these tests are administered to assure a competency level which would allow a student to receive a high school diploma, an actual certificate. In other situations, these sorts of tests are used to permit promotion from one grade to another, and to counter the concern that students are being promoted based on their age and other social reasons rather than on their ability. Approximately 40 states have competency tests under development or in place. These tests sometimes are locally developed by school districts to reflect needs of the community and in other cases are developed by the state, where a single test is used as a standard for all students in that state. Accompanying these tests are expectations that students, if they do not perform well, will receive appropriate remediation.

Competency tests have been criticized on a number of grounds. First, there is concern that a single test is not a sufficient basis for deciding on something as important as whether a student graduates from high school. All tests make mistakes about students. The error in testing sometimes comes from the way the test items themselves are developed, the way a student might feel on a given day, or conditions under which the test is administered. In addition to error, competency tests have been criticized because they appear to shift the entire burden for performance clearly to the shoulders of the student, rather than requiring schools to share the burden. On an individual, student-by-student basis, student responsibility for performance is important; but it should be clear that if great numbers fail carefully developed tests, it will not be long before the public discerns that some of the responsibility for failure belongs to the instructional programs in the schools.

Since a central point in discussing the purposes of testing is to give the public an understanding of what is going on in the schools, some educators strongly believe that the tests used to provide such information, and in turn to make important decisions, ought to be publicly available. But there are counter arguments in support of test secrecy, and they revolve around two general concerns. First, if tests are public, will teachers "teach" to them? There have been reports in the past of testing programs whose intentions were compromised by teachers who gave students direct drill and practice on the actual test items appearing on the test. The effect of such practice is to reduce the complexity of a body of curriculum to something that can be readily memorized to pass a test.

Another criticism of public access to tests reflects an economic concern. Tests which were published after administration (to avoid the direct drill and practice potential described above) would require the repetitious development of similar test items. Since test development is not such a quick or easy business, public access to tests would require costly development activity. In addition, much of the control of test development at present is in the hands of private test development companies. Thus, the public financial outlay would be heavy. However, recent experience with errors found after publication suggests the cost may be worthwhile.

Some advocates of public testing take an intermediate view between direct publishing of all test materials and the secrecy with which much of certification testing proceeds at present. Their argument is as follows: if the test is measuring concepts or skills which are truly important, then it should be possible to describe clearly what the test has in it. They argue that publishing the specifications which guide test development and sample items might be sufficient to meet the need for public access. Again, the utility of such specifications would depend upon the importance of the topic and skill being tested; for, as in the case of test items, test specifications are also costly to develop. Thus, the economic concerns would then tend to favor development of test specifications for more important areas, an outcome that critics taking the intermediate position would support. Whether advocating public access to the entire test or to specifications and sample items, some educators argue that there is no other way to assure that teachers and students understand what is expected of them.

Other issues in the competency testing areas relate to the question of how many sources of information should be used to make an important decision. In some states, the competency test has become the single standard which is used for decision purposes such as graduation. In other situations, students who fail the test are given alternative means of securing a diploma (or promotion). In some cases, other sources of information on the student--teacher reports, counselors' judgments, and other performance records--contribute to the decision. Those who argue for using multiple information sources do so on the basis of fairness. Their opponents, however, often counter with the argument that the performance standard set for competency tests is usually very low (typically, 8th grade reading ability suffices for high school competency), and thus anybody with marginal competency ought to be able to achieve at that level.

Tests of Selection

A second form of tests with direct effect on students are tests of selection. Such tests are typically used to address the problem of resource allocation. Assuming that there are only limited spaces in a program or college, for instance, a test might be used to select the most qualified students. Or, perhaps, if special funds were available to remediate students with the poorest reading performance, a selection test might be used to determine who needed help the most. Selection tests differ from certification tests in that the problem they address is "who is best (or worse)?" rather than "what can a particular person do?" Selection tests are comparative, and the results they provide often relate to a "norm" or standard group. These tests are sometimes called norm-referenced tests (NRT). Common selection tests are those administered to evaluate students' admission status for colleges (like the Scholastic Aptitude Test or the College Entrance Examination Boards). Because tests of this sort clearly impact on students' opportunities, many of these measures are being examined for test bias. That is, are identifiable groups of students performing less well on these tests because of reasons thought to be outside the test content, such as differences in cultural perspectives? In later sections, we will explore some of the ideas related to test bias for these and other kinds of tests.

Students' performance on selection tests is usually reported in terms of where a student stands in relation to the average score. On the SAT, for example, the average score for a typical group of students was 500. A score of 550 means the student was above average for the norm group; a score of 700 means the student was well above average. There are tables which allow these scores to be translated into percentiles, so sometimes scores on tests of selection are reported in percentile rank. If a student's score falls in the 88th percentile, this means that 88 percent of the students taking the test scored below that student. A 15th percentile rank means that 15 percent of students taking the test performed less well. It is important to keep in mind that percentile score does not mean 88 (or 15) percent knowledge, or 88 (or 15) percent right answers. These scores are always interpreted in relation to how other people have done on the test.

Even though the purposes of selection tests are supposed to be limited, they often provide general information about how people are doing. For example, there was much public interest in the report of the declining scores on the Scholastic Aptitude Test, where students' average scores dropped consistently below 500, which had been the average in earlier years. While such tests usually consist of broadly based ability items, inferences about school program effectiveness can generally be made. For this reason, a number of study groups were formed to investigate the potential reasons for the drop in SAT scores.

Tests for Placement

In addition to certification tests and selection tests there are also tests of placement. These tests are used to decide where a student will receive educational opportunity most consistent with his/her ability. Some placement tests are diagnostic. They are developed against a specific field of information or set of skills. Students who do not possess given skills are then placed so that they will receive appropriate instruction. This kind of placement test requires careful description of what a student's performance means, so that proper instruction may be made available. Other placement tests are more general. They determine how students are performing in terms of a general average, such as the average reading score of children in the fifth grade. As such, they tend to be less descriptive, and provide only an overall estimate of where a child might profit most from instruction. Supplementing these tests with good diagnostic tests would seem to be a suitable tactic.

Almost all students encounter tests of the three types--certification, selection, and placement--during their school years. In fact, these tests probably represent only a small portion of the testing an individual undergoes as part of school. Intelligence tests, attitude scales, and aptitude measures are also given, and most teachers develop and administer their own informal assessment procedures to find out where a student stands with respect to a particular skill. By asking questions of students or providing them with a small task to perform, the teacher is able to make inferences about their abilities. Teachers frequently devise their own formal tests as well, using multiple choice or other formats. "Tests" may also take the form of essays, projects, and other activities assigned and evaluated by teachers. So it is clear that testing occupies a good deal of attention in school programs.

TEST USE WITH INDIRECT EFFECTS ON STUDENTS

Beyond the tests developed specifically for decisions about individual students, tests are used for a number of other purposes. Test information can be used to make judgments about the quality of an instructional program or the quality of the teaching staff, or to validate the worth of a curriculum or specific practices. When tests are used in these ways, they are providing "evaluation" information to someone, usually an administrator of the program in question. For instance, public agencies often develop new programs to address problems of literacy or other areas of academic performance. They test students and attempt to infer from the test information whether the program is "working," which is often taken to mean that it improves students' achievement.

Many people in the field of evaluation object to the use of test scores as the exclusive basis for concluding that a program does or does not work. They say that we must also find out if the program in question is being used or "implemented" as intended before we can make an inference from test performance. Similarly, although the use of students' test performance to evaluate teaching would seem to be a good idea, comparisons among teachers assume that the teachers are working in similar settings. It is easy to understand how one teacher might have better instructional materials, or students who were better prepared in a previous grade, or a more sympathetic and helpful principal than a colleague does. Differences on any of these, and many other dimensions, would make the use of student test scores to evaluate teaching a difficult proposition to defend. The same, perhaps repetitious, argument can be made about using test results to evaluate a new curriculum. Sometimes materials and procedures are not actually used as planned and thus test performance alone presents an incomplete picture about their effectiveness.

While much more remains to be said, and has been by others, about the various uses of tests, the issues selected for discussion in this booklet are meant to provide a picture of how much testing there actually is in schools. One school administrator recently reported publicly that students in his district spent about 70 hours taking formal tests for certification, placement, and evaluation purposes. This figure did not include teacher made tests and other informal assessments, nor the tests that come as part of many curricula. Devoting 70 hours to formal testing takes on meaning when one compares it with the amount of time devoted to reading instruction, 180 hours, over the entire elementary school year. Obviously, if testing is taking so much of students' time and, by the way, costing as much as it does, educators and parents want to be sure the time is well spent, and that value is received by all participants.

In the next section, we will discuss one type of test which is currently receiving a good deal of attention. Ways of establishing test quality will also be discusssed with an eye to providing a basic introduction to the kinds of questions which should be asked whenever more or different testing is proposed for schools.

CRITERION-REFERENCED TESTS

Criterion-referenced tests (CRTs) are a relatively new type of measure. They are based upon a design which tries to assess how much a student knows or can do, rather than how much better or worse the student is in comparison to other students. Much work on the definition of criterion-referenced tests has been done, but much more is needed. These tests, like other tests, are also subject to error. They only provide an estimate or an approximation of how how much a student knows about a subject--like, for example, American History--because it would be impractical, even if it were possible, to test students on all facets of that subject. The attempt in criterion-referenced tests is to arrive at scores which have meaning with regard to a particular set of information. Various ways have been described to get at what this "particular set" might be. Some people have used educational objectives as the basis of CRTs. They try to develop test items which assess a student's ability, for instance, to solve simultaneous equations. The task sounds simple, but even attempting to assess something as precise as equation solving can become complicated when decisions must be made concerning the difficulty of the material provided, the format of the test items, how the items are put together to form the total test, and so on.

TECHNICAL FEATURES OF TESTS

Test makers have identified certain general problems relating to the development of tests and have adopted techniques to assure the quality of a test. These general problems are sometimes called validity, standards, reliability, and bias.

Test Validity

Test validity is frequently defined by posing a rather simple question: "Does the test measure what it is supposed to measure?" The problems arise when one begins to speculate on how we might go about answering this question. The simplest task would involve looking at test items and seeing if the items seem to measure what they are supposed to. Does a CRT designed to assess subtraction skills require that students solve problems requiring subtraction? The answer to this question may be made fairly easily if the test consists of sets of numbers requiring subtraction computations. But the issue becomes more complex when the test's items consists of word problems. Then the question becomes: How much of a student's score relates to ability to read as opposed to ability to subtract?

Another problem resides in the fallibility of human judgment. How often would a group of experts looking at the same test agree that it does indeed measure what it is supposed to measure? Until recently, techniques for getting at such judgments were very primitive, and consisted of on-off choices: a test item was either satisfactory or it was out. More recent work has attempted to focus the attention of judges on specific aspects of the test items: for example, whether the content covered represents the full range of the content intended, whether the format and directions for the test are appropriate for the students being tested, whether the language level required for performance is appropriate, and whether the intellectual level of the test item is the same as provided in the objective or test specifications.

The problem of validity inspired some work in a sub-species of criterion-referenced testing called domain-referenced testing. Domain-referenced tests (DRTs) are built by following specifications, much as a house is created from a blueprint. Features of these test blueprints vary, depending upon their designers, but they almost all have certain essential characteristics:

1. Items are referenced in terms of specific intellectual level: that is, the test may measure recall of information, application of information, or some higher order problem solving skills. Various strategies have been developed for deciding on intellectual level; the two most popular were developed by Benjamin Bloom and Robert Gagné.
2. Items are referenced in terms of the performance expected: that is, what the student is asked to do on the test--selecting an answer from a set of alternatives, for instance. As it happens, there are some vocal and persuasive critics who think multiple choice tests should not be used at all since they do not represent the reality of most tasks which confront people.
3. Items are usually referenced to the content they are supposed to cover. Thus it can be determined if the items are representative of the full range of information that the student is supposed to have and if they are fair in that they do not focus on small bits of information.
4. Items are subjected to scrutiny in terms of the way "correct" answers are determined: Is there a rule for knowing the correct answer when you see one? How are "wrong" answers, perhaps provided as foils in a multiple choice item, developed? Are there regularities so that students who consistently select a certain kind of wrong answer would provide good information to

> the teacher about needed changes in instruction? In a writing task, for instance, one might ask what criteria are used to determine whether the paper passes. Sentence structure, spelling, and punctuation are all common criteria, but does the specification also intend that coherence, organization, supporting detail, and point of view be included?

Further research will probably generate other dimensions but at the moment, the four previously described seem to represent a precis of the most common components of domain-referenced testing.

Sometimes people establish "domains" empirically. Someone may decide that "reading" means uttering aloud sentences found on the front page of the New York Times. That would be a domain specification, and we certainly would know instantly how to determine whether an item (or sentence) was fair. Others may generate algorithms, or rules, to decide what a domain is. A simple algorithm might be that all words that start with a single consonant and have a consonant-vowel-consonant pattern (like a special favorite, "cat") would fit. These algorithms might be amplified, for instance, to exclude any nonsense words (like "lat") or any words which appear in the child's primer (like "let"). Certainly, many more complex algorithms or rules have been generated.

Among the problems with domain-referenced and, in some respects, with certain criterion-referenced tests, is the issue of homogeneity. Homogeneity simply means sameness, as particles in homogenized milk are thought ideally to be the same. Some proponents think that the difficulty of items from a CRT or DRT should be the same. Item homogeneity can be judged, again using a set of specifications by which to compare items. Item homogeneity can also be determined using empirical methods. Empirical techniques require that student performance samples--their actual responses to items--be collected and analyzed. In the most absolute sense, we would expect students to perform similarly on all test items from a single domain. But just as we know there is variation in the size of particles making up milk, we also know that certain bits of content are likely to be harder than others. The problem 20 + 30 is easier than the problem 78 + 45, even thought the content specification might say "pair of two digit numbers." While it is probably possible to analyze problems so as to produce homogeneity or equal difficulty, the great numbers of domains which would have to be separately prepared makes the task an impractical one. Another issue is the desire for generalization or transfer. We would hope that when students learn to add such problems, they would then be able to transfer their skill to a wide range of similar problems. Some test makers argue therefore for a tradeoff of specificity (and equal difficulty) for transfer.

Other factors must be taken into account in the validity issue. One has to do with the variations that students might have in their instructional backgrounds. Theoretically, we would like the student to have no experience with an item (and receive zero on a test administered before instruction) and great skill with the item (demonstrated by perfect response) following instruction. But people know things in bits and pieces, and someone is always a little more experienced than another in every area. Thus, the perfect case of complete failure and complete perfection rarely holds in test practice. The problem is further complicated because all of us want to be able to "validate" a test against the kind of ideal instruction which should be offered in a particular area. Such validation can never be complete, for how are we to know what "good" instruction is? Despite the fact that quality inferred for instruction and quality inferred for test items are always mixed up, or confounded, people do make approximations, and look for, in general, improvements of a group from pre- to post-instructional testing.

Setting Standards

A related and similarly confusing issue is the setting of standards. A standard in testing means the same thing as it does in common English: a point against which performance is to be measured. Standard setting can be based upon experiential or more rigorous empirical information. In experiential standard setting, educators and others decide upon a criterion for determining the score a student must attain to pass a test--for certification, for example. A commonly used criterion is the 70 percent level of performance. This standard corresponds to what is usually considered a "C" in the grading of public school academic performance; yet in the classroom, a "C" usually connotes an average score rather than a barely passing score. Since convention and habit play an important role in this form of standard setting, some of the implications might be explored. The charge of "arbitrariness" is often leveled at standards of this sort, particularly when students may be adversely affected, e.g., denied a high school diploma. One might argue that 68 percent isn't really all that much different from 70 percent and so the question arises as to whether a student who achieves at that level should also pass. Such arguments are not easy to live with, and the only sensible reply usually relates to the need for some standard plus the conventional approval given the level of choice.

An issue related to this form of standard setting is the possibility of manipulation of performance. An interesting example occurred in some of the early days of programmed instruction where the military adopted a 90 percent student performance standard for the evaluation of training materials. When such a standard was difficult to achieve, the items on the test were sometimes simplified, allowing the performance level to be "magically" met. Having a clear set of test specifications or item preparation rules reduces this type of practice, but most school systems have not developed their materials in such a way that test item manipulation is precluded.

A second kind of performance standard setting derives broadly from empirical information--information about what students actually do. For instance, existing information which reveals that students are presently able to read certain word lists with 75 percent accuracy is used to justify the selection of a standard pegged to current ability. On the one hand, this kind of procedure seems to assure that the system will "maintain" a level of performance, even in the face of different types of entering students. On the other hand, using existing performance as a basis for standard setting can be seen as a way to perpetuate the _status quo_ rather than as a way to encourage renewed effort.

Another empirical procedure that could be used to set student standards is to identify a subgroup known or thought to be performing at a desirable level. In written language, for instance, standards might be set by inspecting the quality of work of students who are entering college or who may have passed some other test presumed to be valid. This "mastery" group's performance can set a standard that is desirable for the larger population. A more precise use of empirically oriented standard setting involves the close examination of an operationally defined successful group. For instance, the Geometry I course performance of students who go on to succeed in Geometry II might be examined. Their level of performance could then serve as the standard for what is minimally acceptable.

It should be noted that most standard setting done now is of the experiential sort, and the techniques developed for use by school people often focus on ways of obtaining more precisely stated opinions of what students should be able to do and how well they should be expected to do it.

Up to this point, we have limited the discussion of standard setting to the area of deciding upon the level an individual student should meet in order to be passed or certified. Again we confront the issue of test error. Tests can misclassify students, and arbitrary standards can increase the potential for misclassification. People can be misclassified in two clear ways. On the one hand, a person who, if the real truth were known, deserves to pass, might fail on the test and thus be wrongly held back. On the other hand, a person who passes a test may, again if truth were available, really deserve to fail. One question is, of course, which kind of error is worse? The "benefit of the doubt" point of view would hold that it is better to err in the direction of leniency. From this standpoint a person who wrongly passes a test will be in the position of having, in a sense, the benefit of chance in his/her future life. The consequence of this sort of error is that these people may encounter frustration and future failure. For instance, of the Geometry I students who pass, those misclassified might experience extreme difficulty in the next course. In the face of such difficulty, they will probably require more assistance from the teacher, and they may or may not be able to work hard enough to catch up. To regard this kind of error as more serious than an error in the other direction, one would have to have confidence in the test used to make the pass-fail decision and the standard that has been set.

We must also consider the student who fails when he/she should have passed and is therefore held back. This kind of error means that the student will be required to undergo instruction which is largely redundant and an unnecessary cost, and he or she may suffer a morale problem which could discourage persistence in an area which should be well within his or her competency.

There are statistical methods, based on test performance of many students, which theoretically permit the setting of standards to reduce one or the other kind of misclassification. Because decisions of this sort are generally reciprocal, it means that if we try to reduce one kind of error, we increase the other. Some educators, therefore, prefer the equal chance that both kinds of error will occur. It is extremely important to remember that not only can individual items "make mistakes" about people, but the kind of standard set can also contribute to errors and misclassification.

Standards may also involve a more complex set of rules than simple pass or fail. There have been cases in recent competency test development where standards have required a minimum average score, say 70 percent on the entire test, and a provision that at least two thirds of the sub-objectives have to be met. Sometimes the arithmetic works out that it is impossible to do one without the other, but there are cases where this further refinement of standards seems to have practical importance. In effect, a joint provision of this sort means that a student may not compensate for failing a certain portion of the test by doing well on another. While complex rules for passing are not necessarily seen as a virtue, the availability of alternatives should be made known, and the rules for passing be set with as much information on technique as possible.

There is, as previously mentioned, considerable discussion about the use of a single measure, and perhaps a single standard, to make decisions that are as complex and important as student certification. Some school districts have adopted policies which mitigate the effect of a single poor test score. Some localities have decided that a "band" should be established around the pass score. Imagine that the pass score has been set at 75 percent. Students who score between 65 percent and 75 percent could be individually reviewed. In this review, teachers' judgments and student performance on other indicators, such as different tests or written assignments, could be studied. A decision could be reached either to pass the student from a review of his/her record as a whole, or to permit the speedy readministration of the test. Such a procedure formally recognizes the limitation of a single test, the error inherent in all tests, and begins to make the process more humane, if not more sensible. The way the "band" around the pass score is set may vary considerably, again ranging from an arbitrary decision to one based on some rather refined statistical procedures. The force of this general kind of decision, however, is to specify a margin of error and to bring some human judgment to a process which is largely quantitative in orientation.

Another kind of standard which is often used in cases where test findings have indirect effect on students is the proportion of students meeting a criterion. For example, in evaluating the success of a school district's reading program, a question which might be raised is "How many students passed... Fifty percent.... Ninety percent?" The decision relating to the sufficiency of the proportion of passing students can also be arbitrary, but it has inherent in it the same expectations as for a single student's passing or failing. That expectation is the belief that below-standard performance imposes a requirement for remediation. In the case of a single student, clear instructional remedy should be attempted to bring a student's performance up to par. In the case of a system, the expectation is that the school district will continue to invest its time and resources in refining the program until an established proportion of student success is achieved.

An illustration from the field of curriculum development might expand the notion. Suppose it were decided, through some approved but mysterious process, perhaps even through a political process, that a curriculum should be revised by developers until 80 percent of the students attempting it are successful. If on the first tryout of the materials only 60 percent succeeded, the developers would make inferences from performance and try again. But then suppose, after an extended period of time, tryouts, and expenditure of resources, it became clear that under the present conditions the curriculum would never do more than reach 75 percent of the students successfully. Should the developers keep trying? Should they give up? Should the materials be discarded? In the case of a school district which did not meet such a standard, should teachers be given in-staff development experiences? Should the superintendent be thrown out? Should a new school board be re-elected? Some of these alternatives are rather extreme, but are provided to underscore the point that standards create expections, and expections, when not met, certainly cause trouble.

One alternative to the selection of proportion-of-group standards requires that an empirical base be established. For instance, how well do the best schools in a region do when other factors such as wealth and educational levels are taken into account? Maybe the proportions achieved in those settings might be adopted and feasibly met. Perhaps, on the other hand, a broader based information set might be inspected. Suppose information were kept over time about the percentage of increase or decrease in the performance of the school district, state, or region. Tagging performance standards to trends of this sort might also allow them to be more practical and, at the same time, avoid the creation of unreasonable expectations. Scholars in educational measurement have attempted studies in which various procedures are used to set standards, but the political and practical consequences of such processes must always be kept in mind.

Test Reliability

Test reliability is another concern which must be addressed in any discussion of the problems of test use. In the testing context, reliability means the same as it does in common English usage: consistency. Test error, either within the items, on the whole test, or with regard to a standard, depends upon the reliability of the test and how this feature is estimated. At a minimum, we would want a test to measure a particular concept of competency consistently. This means that items on the test should relate closely to one another and consistently assess the student's competency. We would also want the test to be a stable indicator of performance. For instance, a test would not be any good to us if a student had a good chance of receiving a high score on it on one occasion and a very low score on another. We would expect that people, more or less, would perform consistently on the test from one time to another. There are numerous ways of estimating a test's reliability, and most are based upon whether the test differentiates between high and low performers. Reliability can often be increased simply by adding more items to the test. Thus, the longer the test, in general, the more reliable it is. On the other hand, there certainly is an upper limit in practical terms of test length because of concerns for administration time (one or two class sessions) and fatigue of students who may tune out or just get too tired to give accurate responses.

Reliability is often described in terms of a number. A test which has test items that are perfectly consistent with one another, or a test which measures people from test occasion to test occasion with perfect stability, would have a number of +1; the worse case, complete reversal (the highest scorer on one occasion is the lowest scorer on the second) would have a number of -1. Most published tests have reliabilities of around +.8 or .9.

One of the problems with the standard reliability coefficient is that the statistical formulae developed to estimate it are most appropriate for norm-referenced or selection tests. Reliability coefficients suffer when there is little spread among test scores. From the discussion of CRTs, we recall that it is possible on those tests for people to perform quite well if the instruction is successful. In an ideal case, the scores would not be spread out but instead would be clustered at the high end of the scale. Thus, applying conventional reliability techniques to a CRT would provide information which is potentially misleading. Researchers have been exploring the creation of other ways to assess reliability for CRTs and newer forms of testing. Some people have called for procedures which do not depend upon statistics at all, but rather require the test developers to do student-by-student studies without summarizing and aggregating performance over many students. The problem of reliability has also been assessed in terms of setting different standards of performance, but as yet no generally agreed upon solution has been found. Thus, while the concept of reliability is still very important, the utility of such "good" numbers for CRTs and DRTs is still under study.

Test Bias

Test bias is a provocative concept. Bias is often described as something which should be avoided at all costs, and a test should be purged of its baleful consequences. Bias literally means the tendency for the test to give results which systematically deviate in a given direction. A test would be biased if its results were always much worse or much better than what we expect, given some "true" estimate of student ability. Bias of this sort is not terribly important to test developers and users because all persons who take the test are affected in similar ways. But bias does became a serious issue when a test seems to provide differential results for different groups of people. If a certain group of people perform systematically worse on a test than other people do, the test would be thought of as biased against members of the lowest scoring group. For example, if a group of children who have been instructed on concepts of energy do less well than should be expected on a test, a test bias explanation is that there is something _in the way the test itself was constructed_ that leads to the differential results for the group in question. The idea of bias is related to the idea of fairness. Tests, if they are valid, should measure fairly the concepts and skills in question rather than "other", perhaps irrelevant, behaviors.

But test bias remains a touchy issue. For instance, if a group of people regularly scores lower on a particular instrument, what are the possible explanations? As indicated, first, there may be something wrong with the test. The test might use examples that are not common in the experience of the group. The test might have items with especially difficult syntax for group members. The format of the items might be one with which the group had little previous experience. Particular distractors (wrong choices on a multiple choice test) might seem to be "right" because of a certain cultural interpretation brought to the words used. For example, the word "bad" can mean "especially good" in colloquial use among particular groups of people. These factors, and others like them, represent non-essential features of test items that might contribute to test bias. Other features of the test which might contribute to bias relate to the use of sexist and racial stereotyping.

What other reasons are there for biased results in achievement testing? One clear contender is the quality of instruction which students have received. Significant performance differences may occur, not because of problems with the test, but rather because students were not given an adequate chance to learn the material. This line of reasoning is concerned with equity and depends upon a belief that certain groups of students--minority students, as an example--do not receive equal <u>instructional</u> opportunity. In this case, equity means that for <u>all</u> students the actual delivery of instruction, including vague matters such as the teacher's belief that students will profit and learn, is comparable to that received by good performers on the test. Obviously, the extension of bias to include not only features of tests but characteristics of instruction requires a broader range of ideas about how such problems might be solved.

Another issue in test bias, and a problem with all formal testing, is the problem of inferring learning from test performance. Because a child does not perform well on a test does not necessarily mean that good performance on the test is beyond the student's ability. Students may not comply with testing requirements because of overall disaffection with school and its routine. Thus, biased results on tests may relate to the willingess of the student to play the game, to respond on cue and on time in a testing situation. That decision, to play the game or not, is a complex one and in part is based upon the student's estimate that taking the test is worthwhile, that tests previously taken have been pleasant experiences, and that the authority imbued in the test should be obeyed.

Another possible explanation for test bias is that students in low performing groups are not able to learn the material because of individual differences in ability. While a few educators have forwarded this position, the evidence in its support is not compelling. This view assumes a "can't do anything about it" frame of mind, a concept difficult to justify ethically. Furthermore, until the relationship among good teaching, good learning, and test performance is untangled a little more, our efforts should be directed to improving both the design of tests and the corresponding instruction to meet goals.

Returning to the test itself, how are biasing influences avoided? Under the pressure of groups advocating equity for women and minorities, test review procedures have been adopted in which test items are inspected for possible biasing features. For example, in story problems there have been efforts to require that roles and employment options be equally distributed among minorities, women, and white men. Thus, in some tests, and texts as well, more women and minorities are shown in technical and professional jobs, in power positions, and doing more interesting kinds of things than ever before. The linguistic and cultural analyses to which most tests have been subjected are cursory indeed. Often, the only thing done to assure "equity", beyond reading to detect any obvious slurs in the test items, is the imposition of a readability formula to calibrate the test to a certain "grade level." However, most of the techniques in use were not developed for the short, staccato form of writing exhibited by test items; and thus assurances, even on this general level, are weak.

Work in test bias continues to go forward. However, it will probably only have significant impact when the review processess for anticipating bias are made more rigourous and the instructional contribution to test bias made more explicit.

TEST USE

So far, we have examined the major purposes of tests, some different types of tests, a particularly promising alternative to common tests-the criterion-referenced test--and features of tests such as validity, standards, reliability, and bias. There is, of course, more to come. When one confronts the enormous amount of research effort which has been invested in studying tests, and at the same time imagines the time and resources which go into the testing activity, a logical question follows: Are tests useful?

Tests may be useful in two distinct ways. They may have political utility and educational improvement uses. First of all, tests allow school districts and other agencies to look efficient and accountable. Giving tests seems to be a responsible way to act; schools giving no tests would be under suspicion (what are they trying to hide?). This "responsible image" use of tests pertains to other government agencies as well as schools, for tests are often used as a basis of evaluation for innovative programs. So one major use of tests is the appearance function.

Appearance must always be linked in one way or another with reality. What is the reality of test use for the actual improvement of education? To start, no one would make the claim that test results should be the single most important basis for decision making, either on a policy level or in the relative privacy, if not quiet, of a classroom. But the evidence so far does not even show test results as an important feature in instructional decision making. While analysts believe tests should be helpful, the single most important finding in this area is that test results are not now useful. Teachers report that tests do not help them make decisions about students or, more importantly perhaps, decisions about how they would address instruction. Tests seem to be regarded as an administrative burden, something that has to be "gone through" rather than occasions which have promise for application to teaching. Sometimes test data are processed and large amounts of effort go into reporting the results so that they may be easily interpreted by teachers. To date, however, there are precious few examples of school districts which have in place testing systems that truly contribute to the improvement of instruction. Just as an aside, so that it is clear that no shot is being taken at teachers, the use of test results for administrative decision making is also weak. So we are in an interesting and complex position: First, tests are being administered with increasing frequency, and their importance for individual students is on the upswing. Second, the quality of tests available needs significant improvement. Third, there is little evidence, at present, that much sense is made of tests by teachers and other potential information users.

Our technological selves hurry with explanations:

1. New forms of testing (CRT/DRT) have not been in place long enough to have an impact. (Potentially true.)
2. Teachers do not have much experience with or understanding of the test development process or even how to interpret tests. (Our research tells us this is also true.)
3. The public wants tests. (True, for the present.)

Therefore, and the leap to the therefore is long, we should develop procedures to make tests more useful for people. We might argue that we should bring to bear all that we know about learning and persuasion to influence teachers and administrators to know how and to want to use tests. We believe with training and sufficient incentives we could do the job; we might conclude that it has not been very extensively tried before. (Also true.)

Our more cynical selves might counter argue as follows: If tests are expensive and not useful, why have them? Why not develop alternative ways of assessing the goodness of educational programs and services? Why should we continue to persist in engineering mechanisms to support testing in the classroom? Why don't we admit that the classroom may be the place where formal testing is least necessary, given the range of informal assessments teachers daily make about their students? Why not give it up and let the results of tests, given only for ceremonial purposes, comfortably remain in the desk drawers. Let the political use of tests persist. We have better things to spend our time and money on. We are presently trading scarce resources for testing which seems predominantly to have an image building rather than an educational function. Why not build images in other ways? We may have to. But for the time being, the circus is in town and our job is to make it safe for students.

Pouting about tests is not likely to be an effective short-run solution. Instead, we might try a limited program (say about 10 years) to improve the quality of tests and to explore, with open minds, whether they should be used for the purposes we decided upon in the past. This program could be based on four principles which should guide the use of tests in schools.

First, tests, or their specifications, should be public. All people have the right to know what tests are about, how performance will be judged, and how results will be used. They can use this knowledge to prepare for what is expected (should a high school diploma mean 7th grade reading ability?) and to examine the connections among the curriculum, teaching, and testing to make sure that the schools are providing adequate experiences for all students.

Second, tests should be economical. They should be easy and practical to give, reasonably easy to score and interpret, and thus conserving of both money and time. Exploring the cheapest rather than the most comprehensive ways to use tests seems to be an appropriate tactic in times of dwindling resources.

Third, tests should <u>relate to instruction</u> closely and comfortably. Their public nature may help in that process. We should take steps to review the relationship of testing to teaching and teaching to learning to be sure that all the connections are intact. We should continue to explore new ways to integrate testing validly into instruction and attempt to drop the barriers and the trappings which may keep tests from being as useful as they might be.

Fourth, we should be sure that the <u>tests themselves should offer significant experiences</u> to students so that what is tested is important, that the message carried by the tests corresponds to what we think is important in schools, and that the test, in all its parts, is accurate.

These four principles, if used as a basis for the review of a testing process, could result in significant changes in the way people think about tests and believe in their usefulness.

ACTIVISM AND TESTING

This brief section is directed to the activist readers, those who wish to do something about testing and to assure that any increase in testing is worthy of the time and effort expended. Latent activists may also find the strategy presented useful. What inhibits clear public discussion of testing is the perception that tests are arcane and mysterious entities, like the quarks in physics, and that only those who have been initiated are allowed to discuss the topic freely and without reproach. The research and technology which surround testing are sometimes, it is true, difficult to understand, since their language is often equations and derivations, coefficients and calculations. However, common sense questions should be appropriately directed to the testing process, and the belief that only the technically competent are permitted to raise such questions should be eradicated.

Thus, a strategy is proposed which might help teachers, school administrators, and parents begin to pierce the technological armor which has protected tests and their developers. The secret to using this strategy is not to be put off or satisfied with scholarly answers not conveyed in common language. Some good questions are listed below. This set might be useful under the following conditions: (a) when a new test or testing program is proposed; (b) when test results are used to justify new action; (c) when the individual opportunities of students and teachers are constrained by test results. Here are the questions:

1. Why do we need this test? What information will it provide that we don't already have available?
2. How was this test developed? Has it ever been shown that repeated use of this test improved education (teaching and learning)? What kinds of students participated in the development of this test?
3. How much will this test cost, both in money and time taken from other important activities? Is using this test the best way to spend our resources?
4. How are test results reported? Is there a way provided to translate findings into practical courses of action for students and teachers?
5. How much can we know about what is in this test?
6. How do we know the test matches the curriculum?
7. Is there is provision for discontinuing this test if it "doesn't work out"? How will we know it isn't a good test?

Answers to such questions may be hard to get. If these answers are built exclusively on numbers ("the reliability coefficient is .95") or on authority claims ("but famous person X said it was a good test"), you may be suspicious about the reasonableness of the position. But it is clearly important to ask questions such as those suggested so that more tests are not piled on top of existing testing requirements. Unless there is clear evidence that the newly proposed test will improve the situation for any (and preferably all) of the criteria discussed earlier--public accessibility, economy, instructional sensitivity, and significance--one should question directly and listen hard to the answers. The more people within a group, a group like teachers, and the more groups, like administrators, parents, students, counselors, and teachers, which raise the issue of proliferation of testing in an open and careful manner, the sooner the educational community may be in a position to reassure itself about what tests are really for.

SELECTED BIBLIOGRAPHY

Competency Testing

Anderson, B. D., & Lesser, P. The costs of legislated minimum competency requirements. *Phi Delta Kappan*, 1978, *59*, 606-608.

Brickell, H. M. Seven key notes on minimum competency testing. *Phi Delta Kappan*, 1978, *59*, 589-591.

Brickell, H. M. *Let's talk about minimum competency testing*. Denver, CO: Educational Commission of the States, 1978.

Fisher, D. L. *Functional literacy and the schools*. Washington, DC: National Institute of Education, 1978.

Fisher, T. H. Florida's approach to competency testing. *Phi Delta Kappan*, 1978, *59*, 599-601.

Glass, G. V Minimum competence and incompetence in Florida. *Phi Delta Kappan*, 1978, *59*, 602-604.

Hart, G. K. The California Pupil Proficiency Law as viewed by its author. *Phi Delta Kappan*, 1978, *59*, 592-595.

Miller, B. S. (Ed.). *Minimum competency testing: A report of four regional conferences*. St. Louis, MO: CEMREL, Inc., 1978.

Pipho, C. *Update VII: Minimal competency testing*. Denver, CO: Education Commission of the States, 1977.

Pipho, C. Minimum competency testing in 1978: A look at state standards. *Phi Delta Kappan*, 1978, *59*, 585-588.

Wise, A. E. Minimum competency testing: Another case of hyper-rationalization. *Phi Delta Kappan*, 1978, *59*, 596-598.

Test Design

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, *42*, 145-170.

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvements. *Educational Technology*, 1974, 10-21.

Block, J. H. Criterion-referenced measurements: Potential. *School Review*, 1971, *69*, 289-298.

Bormuth, J. R. *On a theory of achievement test items*. Chicago, IL: University of Chicago Press, 1970.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research*, 1974, *44*, 371-400.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 6, 5-10.

Millman, G. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan Publishing, 1974.

Technical Issues

Ebel, R. L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. Washington, DC: American Council on Education, 1971, 625-670.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-162.

Shavelson, R. J., Block, J. H., & Ravitch, M. M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-158.

Overviews of Testing

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

Cronbach, L. J., & Suppes, P. (Eds.). Research for tomorrow's schools--Disciplined inquiry for education. Report of the Committee on Educational Research of the National Academy of Education. London: MacMillan, Callien MacMillan, Limited, 1969.

Levine, M. The academic achievement test: Its historical context and social functions. American Psychologist, 1976, 228-238.

Sanders, J. R., & Murray, S. Alternatives for achievement testing. Educational Technology, 1976, 17-23.

Test Use

Goslin, D. A. The use of standardized tests in American secondary schools and their impact on students, teachers and administrators. New York: Russell Sage Foundation, 1965.

Yeh, J. P. Test use in schools. CSE Technical Report Series. Los Angeles: Center for the Study of Evaluation, University of California, 1978.

GLOSSARY

Achievement test -- used to indicate student knowledge or skills in a particular subject area, e.g., mathematics. An achievement test may be related to specific curricula or instruction; or it may attempt to assess some general level of achievement, e.g., state mandated achievement tests.

Aptitude test -- used to predict or anticipate student potential or capacity for successful performance or learning. Often this is done by assessing students in important component skills or related skills, e.g., verbal fluency tests as predictors of college success, used for entrance screening.

Competency test -- a type of achievement test with some predetermined definition of what constitutes "competent" performance or with some predetermined standard of what is acceptable as competent performance. This standard or definition may reflect the least allowable or lowest limit of acceptable performance. In such cases, the test is often referred to as a minimum competency test.

Constructed item -- a test item for which the student must perform or construct his/her own answer, rather than select one from given choices. An essay test, an oral language test, a driving or cooking test may be constructed item type tests.

Criterion-referenced test -- used to describe student performance in terms of specific "criterion behaviors," i.e., tasks that are considered to constitute achievement in a particular subject area. Criterion-referenced tests are constructed from some sense of what behaviors or skills make up a subject area. For some test developers, this may mean using behavioral objectives to construct the test; for other test developers, a more specific blueprint of skills or behaviors.

Diagnostic-prescriptive test -- used to describe areas in which student performance is inadequate or weak and to suggest subject areas and/or methods of instruction likely to remedy the problem.

Domain-referenced test -- this phrase is often used interchangeably with criterion-referenced test; however, a domain-referenced test is more specifically tied to the subject area (domain) by more detailed definitions of behaviors and skills of the domain and the conditions under which they are performed. These detailed definitions, called domain specifications, provide guidelines that suggest tasks or test items. A sample of domain-referenced items, i.e., a domain-referenced test, is taken as evidence of student ability in the domain or subject area. Furthermore, these domains may be made public.

Item difficulty -- is determined by the percentage of individuals who get an item right. If ninety percent of the examinees were to answer an item correctly, the item would be easy. Conversely, if only ten percent of the examinees were able to answer it correctly, the item would be difficult.

Norm-referenced test -- used for making judgments of relative achievement or relative worth of a student, class, school, or district by comparing that one score against the distribution of scores for some larger group of comparable students, classes, schools, or districts. Comparison results are often reported as percentiles or stanines.

Reliability -- refers to the consistency of test results; that is, to what extent a student's test score varies due to chance or test error.

Equivalent forms reliability is established by giving two forms (equivalent or parallel) of a test to the same person and determining the consistency or agreement of the results.

Internal consistency reliability occurs when most items on a test measure essentially the same thing. It consists of high correlation of scores on the different items within the test.

Test-retest reliability occurs when the same test is readministered to the same students after a time interval and produces consistent results.

Validity -- refers to how well a test accomplishes its aim; that is, the extent to which a test truly measures what it claims to measure. This quality is crucial to all tests.