DOCUMENT RESUME

ED 457 244 TM 033 371

AUTHOR	Miller-Whitehead, Marie
TITLE	Practical Considerations in the Measurement of Student Achievement.
PUB DATE	2001-00-00
NOTE	8p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Achievement Tests; *Criterion Referenced Tests; Elementary Secondary Education; Evaluation Methods; *Norm Referenced Tests; *Performance Based Assessment; Program Evaluation; School Districts; Schools; *Test Use

ABSTRACT

Those who are charged with the evaluation of an educational program or school or effects of teaching often must consider how well students do on the variety of standardized tests they take in the course of a school year. This overview describes common types of tests students take in large-scale programs. Criterion-referenced tests are tests that measure knowledge or skills that students should have mastered. They are based on grade-level curriculum guidelines. Norm-referenced testing is quite different in concept. Test developers administer a pilot test to a representative group of students, compute average scores, and then compare student achievement on the tests to how well the average student performed on the pilot test. Performance based tests are usually paper and pencil tests that are not multiple choice. Such tests are used to assess writing skills, computer skills, or skills in a field such as the performing arts. Using test results and other assessments, the comparison of schools and school districts can be problematic. For this reason, the idea of "value added" or gain score measurement has been adopted by some tests to measure a student against his or her own previous performance. This overview of testing and test concepts may help parents, teachers, and community members in discussions of the various tests administered in a local school district. (Contains 15 references.) (SLD)

Practical Considerations in the Measurement of Student Achievement:

Marie Miller-Whitehead

TVEE.ORG

Marie Miller-Whitehead
Director
Tennessee Valley Educators for Excellence
TVEE.ORG
PO Box 2882
Muscle Shoals, AL 35662

There is an almost universal acknowledgment that schools do make a difference. To believe otherwise is to take so pessimistic a view of the educational process as to be almost unconscionable. We have progressed beyond the Elizabethan belief that a person does not (and should not) rise above the station or role in life to which he or she was born. That said, there remains the task of communicating to students, parents, and the general public sufficient information about student learning to aid them in making decisions about how they and their schools are doing. There are many ways that both parents and the public look at a school's effect on student achievement: the win-loss record of athletic teams, the number of active clubs, involved parent-teacher organizations, successful fund-raisers, an award-winning music program, number of AP classes, number of students on honor roll (all-A students), discipline, promotion-retention, scholarships, and so on (McLean, Snyder, & Lawrence, 1998). All of these are quite important to a student's success in school and the school which fails to provide these may open itself up to public criticism (and deservedly so) no matter how well its students achieve on standardized tests.

However, those who are charged with the evaluation of program, school, or teacher effects often must consider how well students do on the variety of standardized tests which students take during the course of the school year. These may be criterion-referenced, norm-referenced, or performance-based (i.e., writing and the like). Most states have adopted or passed accountability legislation which mandates the type of indicators that will be used to evaluate or assess student academic achievement.

Criterion-referenced tests are tests that measure knowledge or skills which students should have mastered. In other words, criterion-referenced tests are based on grade level curriculum guidelines. They usually have sections on reading, language, math, science, and social studies. The multiple-choice unit tests most people remember taking in school were more likely than not criterion-referenced tests as they were based on material that was taught for a specific course and grade level. This is a simple enough matter for the classroom teacher, who can construct such a test based on what was taught in class and on personal knowledge of the students in the classes. For those working at the district, state, or national level it is somewhat more problematic. There is no universal curriculum although many states and professional organizations have published standards and learning objectives for subject areas and grade levels. Criterion-referenced tests are based upon the premise that upon completion of the course or prescribed curriculum, all students will meet a certain level of mastery of the material. This is usually considered to be between 75% to 80% correct. Students should know what the criterion are and should be able to study for the test. Thus, while some small percentage of students are exempt, the assumption is that regardless of which school the student attends, his socioeconomic level, ethnicity, ability level, or prior preparation, he or she will respond correctly to at least 75% of the questions on the test. The plus factor of criterion testing is equality in the level of expectations for all students. Schools which teach a large percentage of disadvantaged, at-risk students where many parents may have little or no formal education are expected to do as well as those whose students are from highly educated, affluent families. The minus is that cut points and standards may then be set low to assure that sufficient numbers of students achieve mastery level. There are a variety of statistical methods which may be used to make these decisions (McLean & Lockwood, 1996; Millman (Ed.), 1997; Popham, 1988; Sanders & Horn, 1995). Criterion-referenced testing may beg the question of whether or not some schools have students which are more difficult to teach than others or whether some students learn more quickly than others. Even if all students attain mastery it is not possible to determine if all students were challenged to do their best, especially if many students achieved perfect or near perfect scores. The brightest may have coasted through testing while other students may have been severely stressed, the "high stakes test syndrome"(Linn, 2000; Madaus, 1993; Wiersma & Jurs, 1990). In criterion-referenced testing the student is measured against the test, not against any other student's achievement. If a school which serves a large percentage of at-risk students does not have as many students at mastery as a school which serves an advantaged student population, does it mean that the school and its teachers are less good than the other school, or that the school is providing a poor education for its students?

Norm-referenced testing is quite different in concept. Generally speaking, test makers administer a pilot test to a representative group of students, average scores are computed, and student achievement on subsequent tests is compared to how well the average student performed on the pilot test. There are a variety of statistical methods used to determine if the test is fair and equally difficult for all groups of students (Camilli & Shepard,1994; Linn & Harnisch, 1981). Thus, most students would be expected to have scores in the average range, and students who achieved at the 75% mastery level would have done as well or better than 75% of all students who took the test. In this method the student is measured against how well other students performed, not against the test itself. Students are not usually encouraged to study for norm-referenced tests as they are constructed to measure a broad range of general knowledge in subject areas such as reading, language arts, math, science, and social studies. Whereas the hope and expectation is that all students will achieve at least 75% on a criterion-referenced test and many will obtain perfect scores, that is not the case with norm-referenced tests. Few if any students are expected to achieve perfect scores on norm-referenced tests. The advantage of the norm-referenced test is that it is possible for parents and counselors to have some idea of how well a student is achieving compared to other students who took the test and thus aid in making career or college decisions. As with the criterion-referenced test, if the student population of a school is not representative of the student population on which the test was normed, schools with a larger percentage of at-risk students than the norm group may have lower than average scores while schools with fewer at-risk students may have higher than average scores. Once again, does this mean that the school with lower scores but more at-risk students is less good than the school with higher scores but fewer at-risk students?

Performance-based tests are usually paper and pencil tests, not multiple choice tests such as criterion and norm-referenced tests. Performance-based tests are usually used to assess writing skills, computer skills, or skills in the performing arts. Writing assessments are generally scored according to certain rubrics, such as organization, content, grammar, and the like. Since these tests are most often scored by individuals and not by a machine or computer, it is important that assigned scores be consistent, that is that any given paper would receive the same score no matter who scored it, with the same weight given by each to organization, content, grammar, and the like (Moore & Young, 1997; Reckase, 1997). Many states have mandated performance-based tests as part of their state-wide accountability systems but they are expensive to administer and require more man hours and training to be scored properly than do computer scored multiple choice tests. Performance-based tests are similar to criterion-referenced tests in that the student is usually measured against the test criteria, not against how well other students performed. The plus is that all students are expected to perform to certain pre-specified standards, that they have an avenue to exhibit creativity not provided by the multiple-choice format, and that they can demonstrate as high a level of skill on the assigned task as they wish to or are capable of. The question for evaluation and assessment is whether it is reasonable to expect that students who have parents who may not write well (if at all), who have few books, no musical instruments, and no computers in the home will perform as well as students who have access not just at school but at home to these advantages. Thus, are schools and school districts which serve large percentages of disadvantaged students and have lower than average performance scores providing a "less good" education than more fortunate schools and school districts, and how can that be determined?

Since most school districts are not completely homogeneous, comparisons of schools and school districts can be quite problematic. Is it possible to ascertain in a fair way whether or not a particular school or school system is doing as well as it could with its unique percentages of at-risk or other identified sub-groups of students? To that end the concept of "value-added" or gain score measurement has been adopted by several states such as Tennessee and school districts such as Dallas Public Schools, among others (Millman (Ed.), 1997). Value-added or gain scores can be computed for either multiple-choice or for performance-based tests except that portfolio assessment is more commonly used for performance-based tests. Rather than being compared to a norm group or to a criterion, the student is measured against his or her own prior average achievement gain score. It is expected that, with a well-designed curriculum and all other things being equal, a student will learn the same amount of material at the same rate of speed from one year to the next, and that variation in the average amount of material learned is due to the influence of the school and the teachers. Thus, students who learn at a slower pace than others will be compared to their own previous average year's gain and students who absorb information more quickly are compared to their own previous average year's gain. This method has been rather hotly debated on several grounds: (a) the "ceiling effect" in which students who have topped out on a test cannot achieve a gain score because they are already performing at the highest possible level; (b) the "floor effect" in which students at the lowest end have nowhere to go but up and thus may have misleadingly high gain scores (Slavin, 1992); (c) a within school effect that would not become apparent until a cohort of students changed schools and the group gain scores decreased or increased according to which school they next attended (Sanders et al., 1994); and (d) teaching to the test on a norm-referenced test is contrary to the concept of norm-referenced testing and results in scores that are not comparable to those of the norm group if they did not receive instruction prior to testing (Shepard, 1990). Thus district administrators, parents, and community members should be aware that comparisons of schools and school districts are often fraught with pitfalls, particularly when attempting to compare apples with oranges. For this reason many systems have a variety of accountability mechanisms in place to aid decisionmakers.

While this overview may perhaps be somewhat simplistic, nevertheless parents, teachers, and community members who are not professionals in the area of K-12 student testing may find this information helpful as a starting point in the discussion of the various state, local, and national tests administered by their child's school district.

References

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. (Vol. 4). Thousand Oaks, CA: Sage Publications.

Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29(2), 4-16.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Madaus, G. F. (1993). The distortion of teaching and testing: High stakes testing and instruction. Peabody Journal of Education, (2), 28-45.

McLean, J. E., & Lockwood, R. E. (1996). Why we assess students--and how: The competing measures of student performance. Thousand Oaks, CA: Corwin Press.

McLean, J. E., Snyder, S. W., and Lawrence, F. R. (1998). A school accountability model. (Paper presented at the Annual Meeting of the Mid-South Educational Research Association). (ERIC Document Reproduction Service No. ED 428 440)

Millman, J. (Ed.). (1997). Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Moore, A. D., and Young, S. (1997). Clarifying the blurred image: Estimating inter-rater reliability of performance assessments. (Paper presented at the Annual Meeting of the Northern Rocky Mountain Educational Research Association). (ERIC Document Reproduction Service No. ED 414 319)

Popham, W. J. (1988). Educational evaluation. Englewood Cliffs, NJ: Prentice Hall.

Reckase, M. D. (1997). Statistical test specifications for performance assessments: Is this an oxymoron? (Paper presented at the Annual Meeting of the National Council on Measurement in Education). (ERIC Document Reproduction Service No. ED 410 283)

Sanders, W. L. and Horn, S. P. (1995). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. Educational Policy Analysis Archives, 3(6).

Sanders, W. L., Saxton, A. M., Schneider, J. F., Dearden, B. L., Wright, S. P., & Horn, S. P. (1994). Effects of building change on indicators of student academic growth. Evaluation Perspectives, 4(1), 3, 7.

Shepard, L. A. (1990). "Inflated test score gains": Is it old norms or teaching to the test? Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Slavin, R. E. (1992). Research methods in education. Needham Heights, MA: Allyn and Bacon.

Wiersma, W., & Jurs, S. G. (1990). Educational measurement and testing. Needham Heights, MA: Allyn and Bacon

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033371

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Practical Considerations in the Measurement of Student Achievement	
Author(s): Marie Miller-Whitehead	
Corporate Source: Tennessee Valley Educators for Excellence TVEE.ORG	Publication Date: 10/4/2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
XX		
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Marie Miller-Whitehead	Printed Name/Position/Title: Marie Miller-Whitehead, Director	
Organization/Address: TVEE.ORG PO Box 2882 Muscle Shoals, AL	Telephone: 256-446-5278	FAX:
	E-Mail Address: marie@tvee.org	Date: 10/4/2001

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com