2017

# Academic Testing, Measurement, and Evaluation in the Classroom

Orlando McAllister
*College of New Rochelle*, omcallister@cnr.edu

# Academic Testing, Measurement, and Evaluation in the Classroom

**Dr. Joseph A. King with Orlando McAllister, MS.**

*"To teach without testing is unthinkable"* ---
The Joint Committee of American Association of School Administrators

This paper will attempt to outline some of the fundamentals of test measurement and student evaluation in the classroom. It is intended to help minimize the inevitable bias which enters the process of classroom evaluation. Although the quantitative approach to testing and measurement is emphasized, it is equally important to recognize the value and utility of the qualitative approach to student assessment. Often the two approaches can be combined to provide a holistic assessment of students' performance. It is also important to be cognizant of the fact that each approach may have its own limitations, depending on what is being measured. An important caveat in this paper is that the precision of results of assessment in education and other social sciences do not parallel results in the natural and physical sciences.

Testing, measurement and evaluation are not unique to the classroom. As a matter of fact, individual, social and human progress require the establishment of some kind of criteria of evaluation and measurement. Even throughout human history when the rigors of the scientific method were crude, testing and measuring were conducted in a common sense manner. Today, in our modern world, assessment also continues in tried and tested common sense ways: the motorist who kicks his/her tires to measure air pressure; the mother who places the back of her palm against her child's forehead/throat to test the body temperature; or "the chef who tastes his/her soup to determine what other ingredients are needed" --- are all forms assessment.

Testing also has its oral component which was used thousands of years ago. The *Socratic Method* for instance, with its question and answer approach to evaluating logical consistencies and contradictions in an argument is a form of oral testing. Indeed, it was not until the mid 19th Century with the introduction of widely used paper and pencil tests in the U.S. that oral testing lost its prominence. Many countries today continue to require oral examination for advanced degrees, particularly to defend an academic thesis/dissertation. The College of New Rochelle, School of New Resources requires all Life Arts Projects (a significant element of all six credit courses), to have a mandatory oral component. Oral testing will probably continue to be a significant and integral part of the teaching/learning process. Its value in an adult non-traditional college program cannot be overestimated. It will also continue to be an important part of the non-academic world in which a successful interview, in effect, is a kind of oral test…the grand finale to a successful job search. In a very real sense, it helps prepare students for the all-important job interview. Like all effective assessments, the oral test needs to have validity and reliability. (These test characteristics will be addressed later).

The terms test, measurement and evaluation are used interchangeably in this paper. The concept of "test" is used in a specific sense: It connotes the administration of a given set of questions intended to determine a realistic numeric value of the characteristics measured. For example, it helps to measure and predict a student's ability and skills in mathematics or English. Measurement, on the other hand, connotes a broader concept which measures personal characteristics, using a checklist, grading scales, observation (not test) or any other device to obtain information qualitatively.

The *essay test*, when used in combination with the *objective test* --- is capable of producing unparalled, significant outcomes for effective use in the teaching/learning process. The essay test on its own is particularly useful in measuring a student's synthesizing skills. For example, measuring clarity of expression; substantiating arguments; documenting research outcomes; integrating and cross-fertilizing ideas in papers or responding to questions regarding significant events/issues. Notwithstanding the critical importance of the essay type test, it is this paper's position that the *objective type* test which includes

multiple choice, short answer/open–ended, matching, or right/wrong test --- offers unique advantages by minimizing subjectivity in testing. Indeed it can measure in greater detail a given stratum of knowledge in a single test instrument --- whether in math, natural sciences, physical sciences, or even social sciences. This argument is substantiated by Hopkins (1978) who cites Chauncey and Dubin, (1963):

> **Given two examiners fully trained in the art of achievement testing or building essay examinations and the other objective test, the examiner with the essay tests probably can do a better job of estimating a student's present skills in creative writing, but the examiner who builds objective tests can provide more valid and reliable estimates of just about everything of school achievement (p. 188).**

It is important to note that we never measure or evaluate people. Instead, we measure/evaluate characteristics of people: their scholastic potential, knowledge of subject matter, honesty, perseverance, etc. This should not be confused with evaluating the worth of an individual (Mehrens and Lehmann 1973, p. 7).

Furst has noted (1958) that "there is a definite relationship among instruction, objectives and evaluation. And, the results of evaluation provide feedback of the effectiveness of the teaching experience and ultimately on the attainability of the objectives themselves for each student." (p. 3)

In addition, Mehrens & Lehmann (1973) conclude that there are several ways in which evaluation aides the teacher: (1) knowledge about the student's behavior is provided; (2) the teacher is aided in setting, refining and clarifying realistic goals for each student; (3) teachers are helped in evaluating the degree to which the objectives are achieved; and (4) the teacher is aided in determining, evaluating, and refining instructional techniques – all this is invaluable in creating classroom climate most conducive to learning.

The aforementioned authors (1973) also argue that evaluation aids the student by: (1) communicating the teacher's goals; (2) increasing motivation; (3) encouraging good study habits; and (4) providing feedback that identifies his/her strengths and weaknesses.

It is often argued that testing will always be criticized because students tend to "*psych out*" the teacher and learn what the teacher thinks is important. This very '*fault*' can be an advantage because it aids the teacher in communicating his/her objectives.

No doubt, there are always a few educators who suggest that test results should not be communicated because of possible harm to self-concept. This seems to be faulty reasoning. Moreover, empirical data suggest that objective feedback improves subsequent performance. Stroud (1946) argues: "it is probably not extravagant to say that the contributions made to a student's store of knowledge by taking an examination is as great, minute for minute as any other enterprise he engages in." (p. 76)

It is our professional opinion that all instructors need to be conversant with test construction --- such as familiarity with the basic concepts and terminology, test validity, test reliability, test blueprint/table of specification, test administration, derived scores, grades/interpretation of grades --- and where to find reviews of the published tests.

***Validity*** can best be described as the degree to which a test is capable of achieving certain aims: (1) Making predictions of the individual tested; and (2) Describing the student's skills/abilities. Mehrens and Lehmann further argue that Validity is sometimes defined as truthfulness; that is, does the test measure what it claims? In order for a test to be valid or truthful, it must first of all be reliable.

> Neither validity or reliability is an either/or dichotomy… Note, however, that a measure might be very consistent (reliable) but not accurate (valid). For example, a scale may record weights as too heavy each time. In other words, reliability is a necessary but not sufficient condition for validity (p. 124).

It must be noted that test results are used for different purposes because tests have different *validity indexes*. This is why a test derives its name from the purpose for which it was used. For example, predictor, placement, diagnostic or achievement nomenclature, reflect the objectives of the test.

Various authors use different labels for validity. However, (French & Michael, 1966) delimit only three kinds of validity (Mehrens & Lehmann, 1973):

1. ***Content Validity***--- is the degree to which the test items sample the domain of knowledge to which inferences will be made. Content validity is very significant in achievement tests because it is used to rank students in a group or class. The test items must be weighted in proportion to the time spent on each topic and the behaviors desired to be measured. For example, in an Experience, Learning and Identity test, if 50% of class time is spent on the Life Arts Project, then the latter project must carry 50% of the summative (final) grade. Content validity is determined by each instructor based on inspection of test items (subjective determination). Note: two instructors may teach the same stratum of knowledge, but can have different emphases in teaching course content. A single test, therefore, will not have the same content validity for both instructors. In such cases, test items need to be weighted differently in the test for each instructor's class.

Some writers speak of *Face Validity* which is not applicable to objective tests. *Face Validity* is the characteristic of a test that makes it easily recognizable or associated with a particular content universe.

There is no numerical expression for content validity. However, criterion and construct validity (discussed in the previous paragraph) can be calculated numerically using the *Pearson Product Moment Correlation Coefficient* **{r}** which is the most commonly used procedure in reporting validity. Content and face validity are the most commonly known indexes used in teacher-made achievement tests.

2. ***Criterion-related validity*** --- is the skill of measuring the relationship between test scores and an independent external variable (criterion). Sometimes criterion related validity is broken down into ***concurrent*** and ***predictive validity***. For example, when an old test and a new test are administered to the same group and they yield the same results or when parallel tests (i.e. forms A and B of a test are administered to the same group) and the information about the criterion data are collected together, we have concurrent validity. It is a measure to test the stability of a student's performance on similar tests or two different but similar tests.

   However, when test results and criterion data are collected separately the result is deemed ***predictive validity***, e.g., aptitude test scores are used to predict future performance, graduation rate --- the criterion, is compared to dropout rate. Criterion related validity is a characteristic of mastery test. *Predictive validity* is used in making administrative decisions. Moreover, criterion related validity is used to describe a student's behavior for achievement and placement.

   **Construct Validity** – is "…the degree to which test scores can be accounted for by certain explanatory constructs in psychological theory." If, for example, the test has construct validity, individual scores will vary because of individual trait differences. Construct validity e.g., will predict students who possess the (*construct anxiety*) and who will score differently from less anxious students. The construct anxiety may be assessed by measuring the perspiration over one's upper lip or in the palm of one's hand. Another example of construct validity is the use of a pencil/paper test to measure performance in music or dance (Mahrens & Lehmann, 1973, p. 126).

***RELIABILITY***

Reliability is commonly defined as a degree of consistency between two measures.

> … There are many reasons why a student's test score might vary. The amount of characteristic measure may change across time (*trait instability*); the particular questions asked to infer a student's knowledge could affect his/her score (*sampling error*); any change in directions, timing or among rapport with the test administrator could cause score variability; administrator error; inaccuracies in scoring a test will affect outcomes (*scoring error*); and finally such things as

> health, motivation, degree of fatigue of the candidate, and good or bad luck in guessing could cause score variability (Mehrens & Lehman, 1973, p. 103).

There are many known methods used to estimate reliability. Each differs in how they allow different sources of error to surface. However, the most commonly used methods to estimate reliability include but are not limited to *measures of stability* which is called test/retest reliability. It is obtained by administering a test to a group of students and re-administering the same test to the same group or individuals at a later date. Test scores are then correlated. Any change in score is attributed to measurement error. This form of reliability is used to predict the score a student will receive if the same test instrument is administered at a different time. Note that the estimate of reliability will vary with the length of intervals between test administration. Also, because test can be reactive, students tend to perform differently when they are aware of being assessed.

***MEASURES OF EQUIVALENCE***

Measures of equivalence are obtained by administering parallel test or similar forms of a test the same day and correlating the test scores. Measures of equivalence help us to predict how well a student will perform on a similar test with different items.

***INTERNAL CONSISTENCY***

This may be obtained by using the *split half method of estimation of reliability*. It is somewhat similar to using the equivalence/parallel forms of testing. However, instead of two different test administrations, the scores on odd or even numbered items on the same test are correlated. This is an estimation of the reliability of a test half as long. The instructor can then calculate the reliability of the whole test using *Spearman Brown Prophecy formula*. The significance of split-half reliability permits economies of time and cost instead administering two different tests. It cannot be overemphasized that the instructor needs to be clear about the purpose of a test. As noted previously, tests serve a variety of functions and the purpose for which test results are used determine the most appropriate test. The objective of the test should be written in behavioral terms and must begin with an action verb which describes the behavior to be measured (familiarity with taxonomies of education: Bloom 1956; Gagane 1965 & Krathwol, 1962) are essential.

The taxonomies rank cognitive skills/abilities on a hierarchical scale… from lowest to highest. For example, knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom, 1956). N.B. General education goals lack specificity and do not give clear direction to the teacher. Educational objectives give direction both to the teacher and the student Hopkins (1998) cites (Bloom, 1961).

> Participation of the teaching staff in selecting as well as constructing evaluation instruments has resulted in improved instruments on the one hand, and on the other hand, it has resulted in clarifying the objectives of instruction and making them real and meaningful to teachers… When teachers have actively participated in defining objectives and in selecting or constructing evaluation instruments they tend to return to the learning problems with greater vigor and remarkable creativity.

Again, test items should be weighted in proportion to the time spent on areas of the content universe; or the stratum of knowledge to be tested. A rule of thumb to follow is that in objective tests the longer the test the greater the reliability. Tests used with individuals must have better reliability (*.89*) compared with test used with groups (*.69*).

Note (1): Timed tests measure speed and accuracy. Students are expected to answer correctly all questions attempted but are not expected to earn a perfect score; (2) Power tests provide the opportunity to test students' on a given domain of knowledge; students are given adequate time to attempt all questions but are not expected to earn a perfect score on the test because of the difficulty index; (3) Formative evaluation is administered at intervals to give the teacher feedback on the teaching/learning process, what, if any,

changes need to be effected; (4). Summative evaluation is the final test administered in a course to determine final outcomes/course grade.

Generally, tests must be administered in an area that is comfortable and free from distractions or disturbances (place in notice in full view: testing in progress. **Do not Disturb**).

All test directions should be clear, concise, and specific. In a timed test at given intervals the remaining time must be written in a location clearly visible to all examinees. The basis on which grades are to be assigned must be determined prior to test administration. All grades must be valid, accurately calculated and should reflect demonstrated achievement.

Finally, a word of caution when grading students' work:

The most commonly used grades are *grade equivalent* (G.E.) or percentages. These grades are invariably used because of their familiarity by both parents and instructors. Nevertheless, G.E. are the most commonly misinterpreted e.g. a third grader may receive a G.E. of 8.6 on math. The most common misinterpretation, in many cases, is that the third grader should be placed in an 8th grade math class. In reality, all that the 8.6 G.E. implies was that the third grader did that sample of work (test) with the skill and proficiency of an 8th grader. There is no implication whatsoever that the third grader is *au fait* with the 8th grade math curriculum. Another weakness of G.E. is that it is based on extrapolation and assumes that all students grow mentally and physically at the same rate. This is a fallacy. Furthermore, G.E. is not very useful e.g. a student scores 4.3 in reading an 8.6 in math. No comparison can be made between the performance in math and reading, though it is commonly interpreted to mean that the student is twice as good in math as he/she is in reading. If, however, derived scores e.g. *Z scores* or *Stanines* are used, they provide more useful information about a student's performance/abilities. For instance a score of Stanine 4 in math and a Stanine 8 in reading implies that the student is twice as good in math compared to his/her performance in reading. It should be noted that derived scores are statistical calculations and require special training in academic test and measurement (statistics).

Percentages tell us very little in measuring students' performance/achievement and they can be misleading. If a third grade student is given a score of 50% in English, this is ambiguous -- 50% out of ? What is implied in reality is that the 3rd grader knows 50% of all English that he/she needs to know--- a very misleading statement of assessment outcomes.

Another caution in grading is the marked score or raw score which invariably has "*noise*" or error measurement especially in written assignments due to greater scorer bias. This will be discussed later under scoring the essay type question. Unless we can develop a reliable scoring method or calculate error measurement for a group, the instructor will not be able to determine a student's **true** score. The latter score falls within a range and tends to regress towards the mean score. Perhaps to be fair to a student when grading written assignments a good rule of thumb is to mark on a range in an effort to capture a student's true score e.g. A-/B+ or C/D may compensate for error measurement. Finally, a good rule to follow is to develop grades for a particular test based on local norms, i.e. rate a student in relation to the others in his/her class or group.

## ***ESSAY TYPE TEST***

Over the past 75 years there have been many arguments and counter arguments regarding the superiority of the essay type test compared to the objective type test. However, there is no significant research to substantiate those claims. It is significant to noted that (1) it is relatively easy to construct a 5 to six page question extended response essay test compared to 100 item multiple choice test; (2) it helps to eliminate teacher bias by making them stress the students' ability to supply/select the correct answer; (3) it is the only means teachers have to assess the students' ability to compose, answer and present in effective prose (Ebel & Darwin 1960).

The two most serious disadvantages/limitations of the essay type test are: (1) their very low scorer reliability regarding readers' bias, idiosyncrasies, prejudices, despite the thoroughness with which the test

is constructed or read. Also, the vulnerability of the test to the 'Halo effect' – the reader's perception of the quality of the students' synthesizing ability; (2) essay tests provide for very poor/limited, content sampling i.e. the students' raw score can depend on the quality of essay questions set; (3) the reliability of the essay test suffers because of inadequate sampling content, which restricts the number of essays (topics) set in a single test.

Falls (1928) experiment exposed the fallacy of reader reliability by organizing specified scoring procedures using trained readers and a carefully constructed modeled answer. In this experiment 100 English teachers graded copies of a single essay. The readers were to assign a numerical score and to indicate the student's grade level. The results: grades varied from 60% to 98% and the grade level varied from 5$^{th}$ to junior college! Surely the essay test is not significantly reliable since the raw score depends on who reads the test. For example, some readers are deemed easy markers and invariably inflate scores, others on the other hand, are deemed hard markers tend to deflate scores; finally, there are readers who take no chances and tend to award the median score.

Another problem with essay type tests is that students who are unsure or do not have a clear understanding of what the question asked may write well, but have little to say. Finally, the time taken to score essay type tests can lead to reader fatigue and decreased test reliability.

Despite the limitations of the essay type test they are still very popular and widely used, perhaps because they are "easier" to construct and less time consuming. Empirical data suggest that there are too many instances where essay tests are inappropriate – because they tend to measure low level skills on *Bloom's Taxonomy of Educational Objectives* --- because of the frequent use of extended response time essay question "with virtually" no bounds placed on the student.

Mehrens & Lehmann (1973) noted that with the restricted response essay question the student has a better idea of what is asked of him or her. The authors claim that using the restricted type essay can improve scorer reliability and aid the student to synthesize his/her ideas and express them in a logical coherent fashion. Furthermore, this type of essay is of greatest value for measuring learning outcomes at the comprehension, application and analysis levels (*Bloom's Taxonomy*).

A restricted essay needs to be somewhat specific, for example: 'The President of the United States proposes to reform social security and divert part of the funds to private accounts. In an essay (250-300 words), outline three reasons in support or against the president's proposal.'

Mehrens and Lehmann (1973) maintain that good essay questions are achieved with effort and some general consideration. They suggest the following: (1) More time and thought, and effort in preparing each essay type test (use a test blueprint); (2) the construction of tests must have as its objective the elicitation of a specific type of behavior to be measured. (Use Bloom, Erickson's, Gagne taxonomies); (3) a well-designed essay type test must create parameters within which students function (a trade off is necessary between specificity and space in which the student exhibits synthesizing skills); (4) there must be variation in essay topics and spaced over a period of time. But all students must answer the same questions and especially pay attention to the precise wording of questions; (5) more time should be allotted to teaching essay writing and to taking essay type tests; (6) model answers should be discussed with students with the intention of pointing out strengths and weaknesses. Students should be assisted in diagramming their responses; (7) students should be marked for a single difficulty at a time, except for a final test. Other errors should be corrected if time permits; (8) a scoring key should be prepared in advanced to provide a model answer to the essay test. Essays must be graded anonymously with consistency. The mechanics of expression should be graded separately. Grading and scoring essays should occur one question at a time. Standards should be realistic. Moreover, the instructor must be skillful in using the analytical and global method of grading essays. He/she must be clear as to which method is more appropriate at the given time/circumstances. Contingent on discriminatory factors, tests may be scored on a two point scale, acceptable/unacceptable or on a five point scale: superior/inferior. After all is said and done, test scores should not be used simply to assign a grade. An exercise of this sort has little teaching/learning value. The instructor should always ensure that there is adequate feedback/comments, balancing the positive and the

negative. At the same time, students should not be lulled into a false sense of confidence. It should always be remembered that constructive criticisms have merits in a good grading scheme.

> On the theme of essay testing, French's argument is highly relevant: so, if the psychometrician can encourage testing and further classification of those aspects of writing that objective tests cannot measure, encourage the use of readers who favor grading particular qualities that are desirable to grades and see to it that the students are aware of what they are being graded on, we can enlighten rather than merely disparage the polemic of essay testing (French 1965, p. 596) cited in Mehren & Lehmann1973, p. 283.

***THE OBJECTIVE TEST***

Mehrens & Lehmann (1973) also argue that the objective type test format can be subdivided into two categories (i) the short answer test where students supply the answers, and (ii) multiple choice, true/false and matching test in which students select the answers. Objective type tests were developed to decimate most of the criticism associated with essay testing: (1) poor reliability is associated with short tests test and scorers' fatigue; (2) inadequate content sampling which results in fewer test items; (3) dissuade student from engaging in irrelevant padding/bluffing.

A significant strength of the objective test is that it permits an economical method of obtaining information from students, compared to the essay type questions. Objective tests are less time consuming, hence more questions can be answered in a prescribed test. This leads to greater content sampling in the domain of knowledge; it also improves test validity and reliability. In addition, objective tests can be scored faster and more accurately.

Despite the strengths of objective tests, there are many critics who contend that objective tests do not measure higher cognitive skills. They argue that these tests encourage rote learning and guessing, and they deny measurement of synthesizing skills. Some of these criticisms are overstated indeed many prominent *psychometricians* suggest that objective tests can measure the higher mental processes: understanding, application, analysis, and evaluation. It is significant that teacher-made tests concentrate on important facts and knowledge. Trivia must be avoided at all cost. Test items should be consistent with the test objectives, as well as the purpose of the test. Once again, it can not be overemphasized that test items must be written in clear and unambiguous language. Needless to say, the level of complexity of the test will determine the difficulty index. Interdependent test questions should be avoided at all costs. This is because the ability to answer a question correctly should not be based on knowledge of a previous question.

In addition, all test directions should be clear and concise. Students should be comfortable and reasonably relaxed during the test to avoid unnecessary trauma. Students should be instructed to answer all test questions even if it involves "*educated guessing*" since multiple choice type tests are invariably corrected for guessing. In other words, it is assumed that students are likely to guess.

A good rule of thumb to follow is that tests used in making decisions about individuals should have a reliability quotient of *.85*, compared to tests used with groups which should have a reliability quotient no lower than *.65*. Nevertheless, no major decision about a student or a group of students should be made on the basis of a single test administration. Some educators may argue that tests are not accurate. Be that as it may, any information obtained from a test can prove to be significant.

## References


Bloom, Benjamin S. (1956). Taxonomy of Educational Objectives, Handbook I, the Cognitive Domain. New York: McKay.

Bloom, Benjamin S., J. Thomas Hastings, and George F. Madaus (1971). Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill.

Cronbach, Lee J. (1970). Essentials of Psychological Testing. 3rd ed. New York. Harper & Row.

Ebel, Robert L. (1972). Essentials of Educational Measurement. Englewood, Cliff, NJ.; Prentice Hill.

Ebel, Robert L., Dora Darwin (1960). "Test and Examinations in C.W. Harrison" Ed. Encyclopedia of Educational Research. 3rd ed. New York: McMillan pp. 1502 to 1517.

Falls, James D. (1928). Research in Secondary Education. Kentucky School Journal. Pp. 6, 12-46, Mehrens & Lehmann.

French, J.W., and Associates 1957). Behavioral Books of General Education in High School. New York: Russell Sage.

Furst, Edward J., (1958). Construction of Evaluation Instruments: New York, Longmans.

Gagne, Robert M. (1965). "Educational Objectives and Human Performance." In Hellesum A Mehrens and Irvin J. Lehamnn (1973). Measurement and Evaluation in Education & Psychology. New York: Holt, Reihart and Winston, Inc.

Hopkins, K.D., J.C. Stanley, and Br (1990). Educational and Psychological Measurement and Evaluation. 7th ed. Englewood Cliffs: Prentice Hall.

Joint Committee of American Association of School Administrators, (1962). Testing, Testing, Testing, Testing. Washington, D.C.; Association of School Administrators.

Kratwhol, David R.; Benjamin S. Bloom, and Bert Masia. {1964}. Taxonomy of Educational Objectives, Handbook II, The Effective Domain. New York; McKay.

Mehrens, A, and Lehmann, I.J. (1973). Measurement & Evaluation in Education & Psychology: New York, Holt, Richart and Winston, Inc.

Stroud, James B., (1946) Psychology in Education. New York; McKay.

Stufflebeam Daniel L. et. Al. Educational Evaluation and Decision Making. Bloomington, Ind.: Phi Delta Kappa.