Toward a Performance Assessment with Instructional Relevancy and Technical Adequacy: The Case of Curriculum-Based Measurement

Dong-il Kim*

Department of Education, Seoul National University

Abstract

The purpose of this paper is to review the major developments and alternatives in performance assessment and examine the possibilities of the Curriculum Based Measurement (CBM) model. First, characteristics of performance assessment as a viable alternative to objective multiple choice tests will be described. Next, controversy regarding performance assessment will be reviewed. Then, as a exemplary work to overcome weaknesses of the conventional performance assessment, the CBM model will be introduced. Finally, recommendations for future research of the CBM model will be made and implications of CBM in instructional decision-making will be discussed.

Key words: performance assessment, curriculum-based measurement, technical adequacy, basic skill, alternative assessment

* Corresponding author Tel: +82-2-880-7636
E-mail address: dikimedu@snu.ac.kr

Ⅰ. Introduction

It is commonly accepted that the one of the major purposes of education is to facilitate learning. One basis for determining the success of our educational programs is the quality of measurementofoutcomes in learning. In general, performance is the behavior of an individual that can be directly observed by another individual. When this observed behavior changes from one observation to another, it can be inferred that learning has occurred: however, this learning is not directly observed (Shuell & Lee, 1976). Therefore, when the performance of an individual changes as the result of practice or exposure to various experiences, we infer that learning has taken place.

The relationship between learning and performance in educational settings is important. Conceptually, performance can be less than learning, but performance can never exceed learning. That is, you may perform all that you know, but you can not perform more than you know without luck (Shuell & Lee, 1976). Attention should also be given to choosing a relevant performance. There are times, in fact, when a change in performance does not accurately indicate the amount of learning.

How can we make this performance more representative of what has actually been learned? In assessing performance, two important questions arise. The first question is how the discrepancy between learning and performance can be made as small as possible. The second question is what is the best way to provide feedback to the teacher and the student on whether the instructional objectives have been achieved and how learning is progressing. The recent "performance assessment" movement has tried to answer the above questions (Berk, 1986; Deno, 1991; Harney & Madaus, 1989; Neill, 1989; Shepard, 1989; Shepard, 1991).

The purpose of this paper is to review the major developments in performance assessment. First, characteristics of performance assessment as a viable alternative to objective multiple choice tests will be described. Next, controversy regarding performance assessment will be reviewed. Then, as a exemplary work to overcome weaknesses of the conventional performance assessment, the Curriculum Based Measurement (CBM) model will be introduced. Finally, recommendations for future research of the CBM model will be made and implications of CBM in instructional decision-making will be discussed. The focus of this paper will be on the survey of the alternative assessment with instructional relevancy and technical adequacy. Performance assessment is, however, really not new in the theory and practice of educational measurement. More than forty years ago Cronbach (1960) identified three principal features of assessment: Use of a variety of techniques, primary reliance on observations, and integration of information. In distinguishing the assessment from traditional psychometric measurement, he defined assessment in terms of clinical analysis and prediction of performance. Within this context, he emphasized that the manner in which the data were analyzed for decision making was based more on quasi-artistitc synthesis

than on statistical combination. Berk (1986) suggested that what Cronbach meant by assessment is related to the notion of performance assessment today. Berk contended that the technological advances that have occurred during the intervening years have been evidenced in the shift toward the statistical manipulation of the data and the amalgamation of psychometric measurement with impressionist assessment. The notion of performance assessment has been quite popular in administration or management areas (e.g., personnel evaluation) (Berk, 1986; Priestley, 1982), mechanical job performance appraisal (Priestley, 1982) and teacher evaluation (Stiggins & Bridgeford, 1985). It is relatively recent that performance assessments of student achievement have been considered as viable alternatives for solving educational problems associated with standardized objective tests. In eductional settings, performance assessment is identified with direct/authentic assessment (Shepard, 1991) or teacher-made assessments (Airasian,1991). There are several definitions proposed by different authors. They are as follows:

Assessments in which the teacher observes and makes a judgement about a pupil's skill in carrying out an activity or producing a product are called performance assessments (Airasian, 1991, p. 252).

An attempt (performance assessment) is made to determine whether the student has learned the objective. Performance assessment is an integral part of the ongoing teaching/learning process. It is aimed at providing feedback necessary for learning. (Shuell & Lee, 1974, p. 119).

Performance assessment is the process of gathering data by systematic observation for making decisions about an individual (Berk, 1986, p. ix).

Performance assessment, defined as the observation and rating of student behavior and products in contexts where students actually demonstrate proficiency (Stinggins & Bridgeford, 1985, P. 273).

Performance assessment is composed of several essential components. First, the focus of performance assessment is gathering observable data as process or product (Berk, 1986; Gronlund, 1982). To accomplish this performance assessment uses a variety of instruments and strategies. Second, the data are collected by means of systematic observational methods (Berk, 1986; Stiggins & Bridgeford, 1985; Airasian, 1991). The emphasis is on direct observational techniques rather than on paper-and-pencil tests, such as the various multiple-choice test, although such tests may also be employed in the process. Third, performance assessment is aimed at making decisions (Shuell & Lee, 1976; Berk, 1986; Airasian, 1991). These decisions are instructional objectives,

certification, referral and so on.

Interest in performance assessment has been expressed with enthusiasm by those who are disappointed with the current paradigm in educational measurement. The paradigm regards educational measurement as a means of documenting student achievement by using collections of standardized paper and pencil test items (Stiggins, Conklin & Bridgeford, 1986). The main counterpart of performance assessment in this perspective is large-scale objective test with multiple-choice items.

Proponents of performance assessment have pointed out many problems in large-scale objective tests. These criticisms can be summarized as follows. First, the major concern about objective multiple choice tests is the testing-curriculum-mismatch. Shepard (1989) suggested this mismatch could be broken down into two parts: "negotiation of content" and "narrowing of content". Publishers of standardized tests do content analysis so that test objectives are well matched to widely used textbooks. In order to have a "homogenizing" effect, the breadth and the depth of content coverage are limited (Shepard, 1989, p. 5). The test content coverage is negotiated to ensure the market appeal. The content of the test is narrowed so that it cannot cover the full range of important instructional objectives.

Second, bias in standard tests is also a weakness. Neill and Medina (1989) eloquently blamed the objective test as being unfair to minorities. They identified several characteristics of standardized tests that could negatively bias the scores of minority students and of students from low-income families. These tests tend to reflect the language, culture, or learning style of middle- to upper-class whites. Thus, scores on these tests are measures of race or ethnicity and as much as they are measures of achievement, ability, orskill (Neill and Medina ,1989, p. 691). In addition to these, there is another bias in the procedures used to construct norm tests. Questions that might favor minorities are apt to be excluded because they do not fit the statistical properties of the test. Since the minorities represent a low proportion in the sample as well as a low scoring group, test makers tend to discard those items on which low scorers do well but high scorers do poorly. This method of statistical item analyses can produce a test item on which African-Americans do particularly well but whites do not. These types of items are likely to be deleted.

Third, the poor technical adequacy of standardized tests is another problem (Salvia & Ysseldyke, 1991; Neill & Medina, 1989). Standardized tests are regarded as scientifically developed instruments that objectively and reliably measure students' achievement. In reality, however, even the basic requirements of many commercial tests, especially reliability and validity, are open to question. For example, the type of reliability that is generally measured and reported for standardized tests is internal consistency (Salvia & Ysseldyke, 1991). Sometimes, the reliability of subsections ofthetests is very low so that these tests may be not used for placement. In the case of validity, most objective tests do not go beyond the content validity. Test developers (both commercial and governmental) generally validate the content of a test by asking subject-area experts to

make a qualitative judgment about the relationship between individual test items and the trait or traits that the test seeks to measure (Neill & Medina, 1989).

The last criticism of objective tests is "teaching to the test" (Shepard, 1989, p. 5). Since the public gives standardized test scores great weight, teachers tend to teach to the test. Teaching to the test cheapens instruction and undermines the authenticity of scores as measures of what children really know, because tests are imperfect proxies even for the knowledge domains nominally covered by the tests. Large scale standardized tests tend to corrupt the processes of teaching and learning, often reducing teaching to mere preparation for testing (Haney & Madaus, 1989). Plausibly, teaching to the test devalues the meaning of the test results themselves. By having students practice on remarkably similar items, teachers can improve their test performance, but gains from this type of teaching do not necessarily generalize to independent measures of the same content. Thus, test scores can go up without actual gain in achievement (Shepard, 1989; Shepard, 1991).

Ⅱ. Problematic Nature of Performance Assessment.

Since performance assessment has taken the spotlight as an alternative to reduce the disadvantages of large-scale objective tests, performance assessment proponents advocate the use and implementation as well as the related researches (Harney & Madaus, 1989; Neill, 1989; Shepard, 1989; Shepard, 1991). So far, performance assessment has been limited to the exploratory studies in the academic areas of written expression, math, and reading. There is still some confusion about performance assessment on a large scale basis. These problems should be overcome because, in the assessment of students with disabilities, the requests for adequate measures is at a higher rate than ever before. The problematic nature of performance assessment can be summarized in three perspectives: conceptual ambiguity, technical weakness and practical problems in implementation.

A. Conceptual Ambiguity

To begin with, the boundary of performance assessment is rather broad so that it is ambiguous. There are wide variations in the test format. Some advocates strictly contrast performance assessment with multiple-choice objective tests. Under this tradition (Harney & Madaus, 1989; Neill & Medina, 1989; Shepard, 1991), the problems of the multiple-choice format in objective tests were criticized as well as were other drawbacks. They stated that multiple-choice format tests do not measure a respondent's ability to organize relevant information and present a coherent argument. As an alternative, they prefer the "open-ended" question or essay format scored by professional judgments. The term "authentic assessment" (Shepard, 1991) is frequently used in this context.

Beyond a performance assessment with paper-and-pencil test format, another major trend is measuring achievement via behavior or product rating (Berk, 1986; Fitzpatrick & Morrison, 1971; Pristley, 1982; Stiggins & Bridgeford, 1984; Stiggins & Bridgeford, 1986). These types of performance assessment are simulated performance test, work sample, and identification test. Simulated and work sample performance assessment measures a student's ability to carry out procedures or produce products required in school. Identification tests (sometimes referred to as indirect measures of performance) ask students to identify the tools used in science or industrial education (Gronlund, 1982).

Second, Berk (1986) distinguished performance assessment from performance test. According to Berk, performance assessment is a process, not a test or any single measurement device, which may include any kinds of test formats including multiple-choice items. The focus of this process is the continuous data collection through systematic observation, which is integrated for the purpose of decision making.

On the other hand, a performance test is a test in which performance is demonstrated through directly observable behavior as opposed to a paper-and-pencil written response. However, in many instances, these two terms are used synonymously (Stiggins & Bridgeford, 1985; Stiggins, 1987).

Third, there are also various views about the content of the test since the proponents asserted that stakeholders in education have a right to choose different kinds of assessments for different purposes. The "authentic assessment" school like Shepard (and several state assessment program staffs) concentrates on the "higher order thinking". This "higher order thinking" competence includes metacognitive strategies, principled problem solving, and integrated and structured knowledge (Shepard, 1989). On the contrary, the other group placed a high value on automatized skills. These skills are related to more or less basic academic skills like oral reading or spelling: basic component skills must be automatized so as to be integrated into total performance (Stiggins, 1987; Priestley, 1982; Gronlund, 1982).

B. Weak "Technical Adequacy"

Performance assessments also have major disadvantages in technical adequacy (representativeness, reliability, validity) which are directly related to accountability issues. It can not be overemphasized that assessment should be reliable and valid. It is extremely important in special education that a student with a disability is screened and placed appropriately. In special education, the assessment must be not only instructionally relevant but also technically adequate. For example, in the case of reliability, many authorities recommend that the minimum standard reliability coefficient should be .90 when important educational decision like placement in a special class are to be made for a individual student (Salvia & Ysseldyke, 1991). In most cases, experts in assessment evaluate the tests based on whether the Standards for educational and psychological testing (AERA, APA, NCME, 1985) have been followed. According to the Standards, test users should examine the available written documentation on the validity and reliability of tests for the specific use intended. Performance assessment has clear weaknesses in these measurement issues.

1) Validity.

Performance assessments is sort of valid as long as the definition of validity refers to the extent to which a test measures what it is supposed to measure since most of the performance assessments have face validity. Face validity refers to whether the items on a test appear to represent what the test is supposed to measure (Deno, 1991). This kind of validation is necessary for public acceptance. However, there are several ways of test validation (Salvia & Ysseldyke, 1991). The face validity is not sufficient to replace criterion validity or construct validity. To validate a test of a construct, the test author must rely on indirect evidence and inference. The definition of the construct and the

theory from which the construct is derived allow us to make certain predictions that can be confirmed or disconfirmed. As Salvia and Ysseldyke (1991) stated, one conducts experiments to demonstrate that the test is not a valid measure of a construct. Since continued inability to disconfirm the validity of a test leads to test validation, sufficient researches should be conducted. Yet, most of the current performance assessments tend not to provide a evidence of empirical validation.

2) Reliability.

There are also several threats to reliability in performance assessment. Reliability refers to the degree to which test scores are free from errors of measurement (Standards, 1985). Practically, reliability is generalizing the test results across raters, time and forms: interrater reliability, stability reliability and internal consistency (Crocker & Algina, 1986). Performance assessment usually includes a relatively small number of independent observations, low internal consistency and high subjectivity of the scoring process (Meherens, 1991). Many advocates of performance assessment report very little evidence of reliability of the measure (Berk, 1986; Stiggins, 1987). They do not pay sufficient attention to measurement error due to subjective scorers. Measurement errors reduce the reliability. If random error is too great any perceived relevance of the assessment is illusory because nothing is being measured (Fitzpatrick & Morrison, 1971). Thus, valid inference of the measure is based on the reliability.

The only performance assessment area that has reported much evidence on reliability has been writing assessment. The major evidence reported is a reader reliability coefficient which runs in the low .80s (Mehrens, 1991).This reliability estimate considers only the source of error in the rater. There are two kinds of reliability related to the rater: intrarater reliability and interrater reliability. Intrarater vs. interrater reliability refer to the difference between the correspondence of two or more ratings made by the same rater (intra), versus a correspondence of ratings made by two or more observers (inter) (Smith & Teeter, 1982). An index of intrarater reliability is initially important in the development of any observational procedures or scale. It is essential to reduce large amount of variability within individual raters when the procedure is applied several times on the same performance. While intrarater reliability measures consistency within individual raters, interrater reliability provides a measure of consistency between raters. A measure of interrater reliability is called for when consensus across observers is necessary and is particularly relevant when several observers are used in some manner to rate members of a group or when ratings must conform to some type of standard. Although it is seldom noted, it is obvious from the above distinction that intrarater reliability is a requisite for interrater reliability (Smith & Teeter, 1982).

Mitchell (1979) pointed out the problem of raterreliability. For instance, there might be a high correlation between two observers' records of the duration of a teacher's attention to a particular student, but if one rater's watch ran slower than the other's

watch, they would never agree on the actual duration of the teachers' attention. After all, coefficients that use two scorings of the same test (interrater and intrarater reliability) confound random subject error with differences within and between raters.

C. Practical Problems

If any measurement procedure is widely used in the classroom, it has to meet some criteria. Deno (1985) proposed one set of criteria. These recommendations include:

1. Reliable and valid
2. Simple and efficient
3. Easily understood
4. Inexpensive (p. 221).

In practice, performance assessments designed by teachers do not meet some of the criteria. Classroom assessments are designed and constructed by teachers with little formal training in assessment (Coffman cited in Stiggins, Conklin, & Bridgeford, 1986). Many have had no formal coursework and most have had no inservice training in the subject. In another study, Carter (1984) investigated the test development skills of 310 high school teachers and reported that teachers had great difficulty recognizing items written to measure specific skills. She also reported that teachers felt insecure about their test making capabilities. In their research on teachers' use of performance assessment, Stiggins and Bridgeford (1985) found that over three quarters of the 228 teachers surveyed reported using performance assessment and they described their performance tests to often be used with little attention to assessment quality.

Under these circumstances, performance assessment is hardly adequate. Stiggins (1987) cautiously mentioned the potential sources of inaccurate performance assessment such as poor exercise, few samples of performance, vague criteria, poor rating procedures and poor test conditions (p. 35). To obtain a valid and adequate assessment, the assessment should have carefully prepared performance exercises, clear performance expectations and rating system. However, in the large scale assessment, this procedure is not always feasible. It is costly to get an acceptable level ofreliability in performance assessment. It requires careful selection and extensive training of the raters, precise scoring guideline, and periodic rechecking of rater performance (Mehrens, 1991). Performance assessment would be neither efficient nor inexpensive on these conditions. Without sacrificing the benefits of performance assessment, several approaches have been developed to overcome the problems, especially the problem of public accountability which is an important concern for special education. One of them is the Curriculum Based Measurement (CBM) proposed by Deno and his associates (Deno, 1985; Deno, 1991; Fuchs & Deno, 1991; Shinn. Tindal, & Stein.1988)

Ⅲ. CBM as a Performance Assessment

A. What Is CBM ?

In many curricular areas, students are expected to become fluent in the performance of particular skills. To assess skills, assessors often rely on direct performance. Salvia and Ysseldyke (1991) refer to CBM as an example of "frequent direct performance measures in core achievement areas" (p. 550).

CBM originated from three different sources (Deno, 1991). First, in 1977, the University of Minnesota Institute for Research on Learning Disabilities (IRLD) received funding to study alternative methods of special education decision making with learning disabled students. The research on CBM was based, in part,onearlier work by Deno and Mirkin (1977). Their model, then called Data-Based Program Modification (DBPM), included procedures for generating curriculum-based data on student performance. Through the Minnesota IRLD research program, Deno and his colleagues developed a standardized set of measures (Shinn, 1988). This research was also influenced by "the observational and analytical methodology of applied behavior analysis" and "the techniques and methods of Precision Teaching" (Deno, 1991, p. 10). The last major influence as reported by Deno was the application of conventional test theory to the development of CBM in test construction: standardization, reliability, validity. These came together to make CBM a formal and structured assessment model.

The development of measurement procedures were made as follows (Deno, 1985; Deno, 1991; Fuchs & Deno, 1987). First, the characteristics of the measures were identified.and among the characteristics, technical adequacy and efficiency were emphasized (Deno, 1991). The next step was to find alternative behavior indicators of the basic skill of interest. For reading these included supplying words deleted from text, saying the meanings of words underlined in text, reading aloud form isolated word lists, or reading aloud form text passages (Deno, Mirkin, & Chiang, 1982; Deno, 1985). Then, a series of studies were conducted on various measurement parameters (format, duration, source of stimuli). After considering alternative measurement formats, Deno and his colleagues investigated criterion validities (Deno, 1985; Shinn, Tindal, & Stein, 1988). This method of construction was used to develop systematic measurement procedures for reading, spelling, and written expression.

B. Is CBM a Performance Assessment?

In order to determine whether the CBM model is a performance assessment, it is useful to apply the identified characteristics of performance assessment to the CBM. The main characteristic is assessing the observable behavior or product through a systematic observation for the purpose of making decisions.

There are various submeasures that utilize observable behaviors orproducts in the CBM model for gathering data (i.e. reading passage aloud in reading, spelling words dictated from a specific list and writing a composition with a story starter). As described eariler, the CBM model includes data gathering by systematic observation derived from the techniques of applied behavior analysis. The CBM model improves the data base for making educational decisions for instructional change, referral, program evaluation, and so on. Thus, the CBM model covers essential ingredients of performance assessment.

The CBM model has three important assumptions regarding the systematic observation and decision making: direct measurement, repeated measurement, and time-series analysis (Marston & Magnusson, 1987). The first and primary assumption of the CBM model is that assessment focuses upon direct observation of a student's academic skills and behaviors in the student's current instruction. Lovitt (1967) stated that the validity of assessment can be significantly improved with an emphasis on the direct measurement of academic behaviors of concern. The second essential component in the CBM model is the use of repeated measurement of pupil performance. In addition to increasing reliability, frequent assessment increase the validity of describing student change or progress. As a result of documenting student growth, the system can be used to monitor the effectiveness of educational interventions and progress toward the goals. A third major element of the CBM model is time-series of the data display. While direct and repeated measurement of a pupil's performance are necessary conditions for successful implementation of CBM, they are not sufficient to assure the effectiveness of the model. Essential to the process is the graphing of the academic data, and the analysis of students' learning rates in response to educational intervention.

There are also some unique characteristics of the CBM model. These distinguish CBM from the conventional performance assessments which are specific skill-oriented and not defensible psychometrically. The CBM model is a general outcome measure of standardized procedures with technical adequacy. Deno (1985) effectively described the fundamental asset of the model.

Teachers require simple, (reliable) valid, and efficient procedures that they can use to observe student performance in the curriculum of the school-procedures that function as the "vital signs" of student educational health-so that they can make judgments regarding the effectiveness of their efforts to instruct individual student (P. 230).

C. Distinct Characteristics of the CBM

Three important features of CBM which make CBM different from conventional performance assessments will be reviewed: technical adequacy, scoring procedure, and general outcome measures.

1) Technical adequacy of CBM.

The standardized procedures of CBM have made it possible to conduct various reliability and validity studies. The technical adequacy in the areas of reading, spelling, and written expression was investigated in a series ofstudies (Deno, Marston, & Mirkin, 1982; Deno, Mirkin, & Chiang, 1982; Deno, Mirkin, Lowry, & Kuehnle, 1980).

In reading, as described previously, it was determined that counting the number of words read aloud correctly in 1 minute from either a word list or a passage from the curriculum is a valid measure of a student's reading proficiency. The correlation between the oral reading fluency measures and the reading criterion measure, including decoding and comprehension, range from .73 to .91 with most coefficients in the .80s. These concurrent validity findings have been replicated in other studies (Marston, 1982). Internal consistency, test-retest, and interscorer reliability estimates ranged from .89 to .99 (Marston, 1982; Tindal, Marston, & Deno, 1983).

Reliable and valid measures of spelling were identified by counting the number of words spelled correctly or the number of correct letter sequences (a procedure for counting correct pairs of letters, see White & Haring, 1980), written in response to a dictated word list in a 2-min period. The validity correlations between the CBM measures and criterion measures were very high, ranging from .80 to .96. Internal consistency, test-retest, and interscorer reliability estimates ranged from .86 to .99 (Marston, 1982; Tindal, Marston, & Deno, 1983).

Counting the total number of words or the total number of correctly spelled words written in 3 minutes in response to story starters and topic sentences provided a valid index of writing proficiency and correlated well (.70 or higher) with the criterion measures (Deno, Marston, & Mirkin, 1982; Marston, 1982). Marston and Deno (1981) analyzed the reliability of written expression measures and determined that test-retest reliabilities was .81 to .92 for one day and .62 to .70 for three weeks. Internal consistency coefficients, derived from examining performance at the end of 2, 3, 4, and 5 minutes, ranged from .70 to .99. Coefficients of .90 to .99 were found for interscorer reliability.

2) Scoring procedure.

Another distinction between CBM and conventional performance assessments is scoring procedures. In the case of written expression, scorers are provided with analytical or holistical rating guides in the performance assessment (Stiggins, 1987). The scoring methods are composed of several content criteria and rating systems such as the Likert scale, which sometimes have questionable reliability and validity. As a result of the review of the existing literature and examination of potential procedures, simple and direct measures were used to count the frequencyofwordsreadaloud or written rather than rating the complete product. In sum, these short and efficient procedures in the CBM are reliable as estimated by several reliability methods and valid with respect to the other measures of the same academic content.

3) The CBM model as general outcome measurement.

The CBM model differs from the conventional performance assessment in another important way. This model has a salient feature of measuring general outcome indicators which are (a) the assessment of proficiency on the global outcomes toward which the entire curriculum is directed, and (b) the reliance on a standardized, prescriptive measurement methodology that produces critical indicators of performance (Fuchs & Deno, 1991, p. 493).

Developing the conventional performance assessment starts with clear and specific objectives which identifies the decision to be made from the assessment (Airaison, 1991). The various decisions may include individual diagnosis, grading, grouping, selection, certification, and program evaluation at the specific period of time (Stiggins, 1987). Many forms of performance assessment rely on specific subskill measurement.

Yet, these short-term and specific subskill measurements have some problems when they are applied to measure learningrate. Specific skill oriented measures require a shift in measurement focus each time a skill is mastered. Fuchs and Deno (1991, p. 492) described the difference between specific outcome measurement and general outcome measurement. For instance, under a subskill measurement, a learning rate for mastering blends or vowel teams can be obtained. But, overall progress across blends and vowel teams cannot be described, because (a) different skills are measured at different points in time, and (b) different skills are not of equal difficulty and do not represent equal curriculum units.

From an opposite perspective, Fuchs and Deno (1991) contended that the CBM model focuses on the broader final task. With general outcome measurement, teachers can monitor students' development across a school year without any shifts in measurement. Since general outcome measurement samples material across the curriculum, the difficulty of the tests remains constant across the year. Because of this, the CBM may be less sensitive than specific skill mastery assessment. However, general outcome measurement like the CBM model provides a database sensitive to instructional effects, which can be used effectively for instructional decision making (Fuchs & Deno, 1991).

IA. Evaluation of the Current Studies of CBM and Recommendations for Future Research

CBM has received considerable attention as an alternative method of making educational decisions with students with mild disabilities. There are extensive studies that have been conducted regarding technical adequacy and test construction. This section summarizes the related studies and some recommendations for future research are made. Three topics are reviewed: application of classical test theory on reliability

studies, test validation, and parallel forms.

A. Application of Classical Test Theory in Test Construction.

CBM has been through a series of studies which produced high reliability coefficients for interscorer reliability, test-retest reliability, alternate forms reliability and internal consistency. These concepts of reliability are based on classical test theory. In classical test theory, the observed score variance is partitioned into true score variance and error variance. If tests were perfectly reliable, true score variance would equal observed score variance (Davison, 1989).

Since several error factors exist, classical test theory differentiates between intrarater and interrater indicators, order and agreement indicators, and consistency and stability indicators of reliability. Each reliability estimate considers one source of error either in observer, test, occasions or test forms (Eason, 1989). Yet, the classical true score model has been criticized by emerging alternative approaches.

1) Alternative 1: generalizability theory.

Generalizability theory (G theory) subsumes classical test theory as a special case (Eason, 1989). G theory encompasses the concepts of classical test theory as well as accommodating complex measurement designs. The power of G theory lies in the consideration of multiple sources of error variance simultaneously. Classical test theory is limited to analysis of single sources of error variance (Webb, Rowley, & Shavelson, 1988). G theory looks not at how reliable an instrument is over varying situations but rather how generalizable the results are to a universe. A generalizability coefficient represents the ratio of universe score variance (systematic variance) to observed score variance (Crocker & Algina, 1986).

More precisely, the comparable concept of true score in G theory is an examinee's universe score (Brennan, 1983). Universe scores defined as the average of the measurement over the universe of generalization. The universe of generalization refers to the universe the researcher wishes to generalize. Instead of assuming, as does classical test theory, that the individual differences constitute the only lawful source of variation in test scores, generalizability theory assumes that there may be a number of sources of variation (Smith & Teeter, 1982). These sources of variation other than individual differences are called facets. Different scorers, alternate test forms, or separate occasions are examples of facets that might be studied. A particular combination of facets makes up the universe to which test scores may be generalized (Webb, Rowley, & Shavelson, 1988).

As an index of reliability, a generalizability coefficient is computed. The generalizability coefficient reflects the partitioning of variance into components that correspond to the facets sampled in the study [universe score variance / expected observed score

variance (universe score variance + appropriate error variance)] (Crocker & Algina, 1986). The coefficient itself combines these components in a ratio that also represents the proportion of variance attributable to individual differences for a particular universe (Mitchell, 1979). One G study can generate several coefficients, each corresponding to different universe of conditions.

2) Alternative 2: item response theory (IRT)

Throughout the history of classical test theory as a model for psychological measurement, some problems become obvious. It is, partly, in response to these recognized inadequacies of classical test theory that Item Response Theory (IRT) was developed. In IRT, the comparable concept of true score is the possible values of Q (theta), the latent trait (or according to Lord, probability of correct response): Examinee performance on a test can be predicted by defining examinee characteristics, referred to as "trait".

Major shortcomings of classical test theory proposed by the IRT proponents are summarized as follows (Weiss & Yoes, 1988): First, test item parameters like item difficulty or item discrimination defined in classical test theory are dependent upon the sample in which the items were administered. For example, if a set of test items were administered to a high-ability group of examinees, the item difficulties would be different than if the same items were administered to a group of examinees of moderate or low ability. IRT resolves this problem of sample dependency by providing item parameters which are invariant; that is, they are not dependent on the ability level of the group upon which the item parameters were developed.

A second major problem concerns the scoring of individuals. Because individuals are scored based on the number of items to which they respond correctly, test scores are dependent on the difficulties of the items used in the test selected. In contrast, IRT provides scores for individual examinees which use information available on the items administered, but which are not dependent on the specific set of items administered.

A third major problem with classical test theory involves the concept of reliability. "True" scores cannot be directly measured, and must be estimated from observed scores. In classical test theory, point estimates of true scores are derived from the "index of reliability", which in turn is based on the reliability coefficient. Because reliability involves the total (i.e., observed) score variance, both the estimate of an individual's true score and its confidence interval are dependent upon the particular sample of examinees involved in the total score distribution. In IRT, precision of estimations can differ for different θ(theta) level

As a test theory based upon a model, IRT has certain assumptions about the data which must be made in order to hold the model. There are three important assumptions in IRT at present: unidimensionality, local independence and shape of item characteristic curve (Hambleton & Swaminathan, 1985).

It is generally assumed that only one ability or trait is necessary to explain examinee test performance. Item response theory models that assume a single latent ability are referred to as unidimensional. Unidimensionality is defined in terms of the statistical dependence among items. Specifically, the requirement for a test to be unidimensional is that the statistical dependence among items can be accounted for by a single latent trait. This means that a test is unidimensional if its items are statistically dependent in the

entire population, and a single latent trait exits such that the items are statistically independent in each subpopulation of examinees whose members are homogeneous with respect of the latent trait (Crocker & Algina, 1986).

There is an assumption equivalent to the assumption of unidimensionality known as the assumption of local independence (Hambleton & Swaminathan, 1985). This assumption states that an examinee's responses to different items in a test are statistically independent. For this assumption to be true, an examinee's performance on one item must not affect, either for better or for worse, his or her responses to any other items in the test. For example, the content of an item must not provide clues to the answers of other test items.

One should note that the assumption of local independence for the case when θ (theta) is unidimensional and the assumption of a unidimensional latent space are equivalent. First, suppose a set of test items measures a common ability. Then, for examinees at a fixed ability level θ(theta), item responses are statistically independent. For fixed ability level θ, if items were not statistically independent, it would imply that some examinees have higher expected test scores than other examinees of same ability level. Consequently, more than one ability would be necessary to account for examinee test performance. This is a clear violation of the original assumption that the items were unidimensional. Second, the assumption of local independence implies that item responses are statistically independent for examinees at a fixed ability level. Therefore, only one ability is necessary to account for the relationship among a set of test items (Crocker & Algina, 1986). The third assumption has to do with the assumed shape of the item characteristic curve, or as it has been more recently called, the item response function (IRF). This assumption involves the particular mathematical form of the item characteristic curve (Weiss & Yoes, 1988). Most of the current IRT models are based on the form of the normal ogive curve.

3) Evaluation of alternatives.

The remaining question is whether classical test theory is appropriate for the CBM model despite the weaknesses. The question can be answered by examining the alternatives.

First, when you want to adopt the G theory for reliability studies, both individual differences among students and the influence of other factors (situation and time) on scores should be really important enough to do extra work. Is it really worth the additional time and energy to use the G theory? If we don't have explicit gains except the complicated computation of generalizability coefficients, nobody may recommend it. In the CBM model, when it is applied to instructional decision making or program evaluation, most of reliability issues would be handled within the classical test theory. If anyone really wanted to conduct an experiment with a complex design, which needed to separate time, situation, and scorer effects, G theory could be an alternative.

Second, IRT is based upon the "strong" assumptions: unidimensionality and local independence. Other measurement theory (classical test theory and generalizability theory) does not necessarily (explicitly) assume those assumptions.

The CBM model does not fit those assumptions. The measures in CBM emphasize skill fluency, "a combination of speed and accuracy" (Shinn, Tindal, & Stein, 1988). This clearly violates the unidimensionality. In addition, in the area of reading, the same words may appear in one passage. This also violates the local independence assumption. The last and major problem of application of IRT is that CBM is a sort of performance assessment which is oriented toward a whole test rather than an individual item. In CBM model, the response of interest is the total number of student's response in given period. Individual item response function can't be generated.

There is no logical basis for ever concluding that the set of assumptions of a model should be met by a data-set. Determining the adequacy with which a test dataset fits a particular set of model assumptions will be useful information to have when choosing a model (Hambleton & Swaminathan, 1985). When the assumptions of a model cannot be met, the model-data fit will often be poor, and so the model will be of questionable value in any application. The classical test theory is based on weak assumptions, that is, the assumption can be met easily by most test datasets as well as CBM. Unless any researchers change the current procedure and pick out other measures which meet "strong" assumptions or requirements, the reliability studies based on classical test theory in the CBM model remain valid.

B. Validation of CBM Measures

The investigation of validity is related to the interpretability of test. We are interested in what trait is measured by the test. Validity is confirmed when the test is measuring the trait which is intended to measure and measure the trait well. Although there are

numerous types of validity, all fall into three main classes: criterion-related validity, content validity, and construct validity (Crocker & Algina, 1986).

The general paradigm for criterion-related validity involves establishing therelationships between scores on the test and criterion. The CBM model has been well established in this validation as described before. Content validity is the inference from a test to the universe or domain. CBM measures are constructed as high-power and high-speed tests, consisting of many items that students complete within a short time interval. Theeffects of this type of testing versus the content validity of the stimulus items remain to be partialed out with respect to which features contribute to the differences in student performance (Shinn, Tindal, & Stein, 1988).

The most important validity is construct validity which is the degree of relationship between the test and the trait. Messick (1980) emphasized the importance of construct validity for test use. He argued that, even for purposes of applied decision making, reliance upon criterion validity or content coverage is not enough. The meaning of the measure (construct validity) must also be comprehended in order to appraise potential social consequences sensibly.

A study on the construct validity of CBM with regard to reading proficiency and reading comprehension at the elementary level was conducted (Shinn, Good, Knutson, Tilly, & Collins, 1991). However, few studies have been done on the relationship between reading fluency and comprehension in special curricular content area. Few studies of construct validity have been done in other area of CBM measures like math, spelling, and written expression. Further studies are needed to investigate validity in these areas.

C. Generating Parallel Forms

SinceCBMinvolves repeated measurements on equivalent forms of the same task across extended periods of time, it is major requirement to generate tasks of the same difficulty. Fuchs and Deno (1991) described clearly thewaytocreatetheCBMmeasures. For example, in the area of spelling, the teacher creates a pool of all words constituting the particular curriculum, which is the year-long curriculum on which she wants to measure spelling improvement. Then, she randomly samples, with replacement over days, certain number of words form the pool.

In practice, when Marston and Magnusson (1987) implemented CBM at a district level, they made the CBM measures through a standardized procedure. In the area of reading, for instance, rather than had teachers randomly select stories or passages form basal reader the Minneapolis team of curriculum specialists developed a package of reading passages to be used by all special education resource teachers. At each level of Holt, thirty passages with at least 200 words, were randomly selected and then typed on separate forms.

By random sampling in the same level from basal readers, the control of difficulty seem

to be gained at least theoretically. However, generating the equivalent forms is an empirical question as well as a theoretical issue. The empirical procedure of examining the equivalent forms are as follows (Crocker & Algina, 1985). When the two tests meet the requirements for parallel tests, it is possible to establish a mathematical link between the correlation between true and observed scores, and the correlation between observed scores on two parallel tests. According to classical true score theory, two tests are defined as parallel when each examinee has the same true score on both forms of the test and the error variances for the two forms are equal. Such tests will, as a consequence, have equal means and equal variances. The reliability coefficient can be defined as the correlation between scores on parallel test forms.

No research has been conducted on the examination of all equivalent forms in reading. Lynn Fuchs (personal communication, October 11, 1991) was concerned about this issue since great variability of performance in students of learningdisabilities might come not only from their learning characteristics but also from differences in difficulties of passages.

A. Concluding Remarks

Few people question the value o f assessment; each acknowledges in one way oranother that the task of designing methods for assessment efficiently and without undesirable consequences is a demanding one. In general, they seem to agree that the development of a truly adequate approach to educational assessment which not only will measure accurately what has been learned but will also provide useful information for future instruction. After the turn of century, the use of nationally normed testing became increasingly common in schools. Additionally, as a result of federal legislation that promoted testing (the Elementary and Secondary Education Act of 1965) and various initiatives focused on reform of education, it seemed that objective testing received a substantial boost (Haney & Madaus, 1989). However, due to several shortcomings, the current objective tests are likely to be narrow and restrictive. They often represent only a small fraction of the assessments that take place in schools and that influence the quality of schooling and student learning. The emerging interest in performance assessment was expressed under this context. It emphasized the vital openness which is sensitive to individual differences and instructional activity. Thus, conventional performance assessments tend to be more general and ad hoc with respect to the stimuli and the procedures used in measurement. Yet, sometimes, these procedures lack objectivity and public accountability, which are necessary for special education. In recognition of these shortcomings, a different approach was made in CBM.

As a general outcome measurement, CBM is standardized regarding what and how to measure, and the prescribed stimuli and procedures remain stable during the

instructional period (Fuchs & Deno, 1991). By the application of traditional psychometrics and quantification of observed performance, CBM tries to obtain two standards: instructional relevancy and technical adequacy in selected academic areas. This is a sort of developmental synthesis of the traditional objective test and conventional performance assessments in the area of instruction for students with mild disabilities.

Problems of education are by no means limited to assessment of achievement. Even the best alternative assessment, used in the most caring way, will not resolve many of these problems alone. Arguments for assessment must be rooted in the search for answers to basic questions: How do weeducate the students? And for what end? The major feature of the CBM model is directly linked to those questions. CBM contributes significantly to providing a valuable database that may be used across various kinds of educational decisions made in developing and evaluating instructional programs.

References

Ackerman, T .A., & Smith, P .L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12(2),* 117-128.

AERA, APA, NCME (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bennett, R. E., Rock, D. A. & Wang, M. W. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28(1),* 77-92.

Berk, R. A. (1986), *Performance assessment: methods and applications.* Baltimore, MD: The Johns Hopkins University Press.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats - it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11(4),* 385-396.

Bracey, G. W. (1989). The $150 million redundancy. *Phi Delta Kappan, 70(9),* 698-702.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: Act Publications.

Carter, K. (1984). Do teachers understand the principles for writing tests? *Journal of Teacher Education, 35(6),* 57-60.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Rinehart and Winston.

Cronbach, L. J. (1960). *Essentials of psychological testing (2nd ed.).* New York: Harper and Row.

Cronbach, L. J., Gleser, G., Nanda, N., & Rajaratnam, N.(1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Davison, M. L. (1989). Generalizability theory. paper presented to the Conference on Measurement Theories and Appications, Tainan, Republic of China

Deno, S. (1985). Curriculum-based Measurement: The emerging alternative. *Exceptional Children, 52(3),* 219-232.

Deno, S. L. (1991). The nature and development of curriculum-based measurement. Unpublished manuscript. University of Minnesota, Minneapolis.

Deno, S. L. & Mirkin, P. (1977). *Data-Based program modification: A manual.* Reston, VA: Council for Exceptional Children.

Deno, S. L., Mirkin, P.K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 18,* 16-26.

Eason, S. (1989). *Why generalizability theory yields better results than classical theory.* Paper presented at the Annual Meeting of the Mid-south Educational Research Association, Little Rock, AR.

Fitzparick, R. and Morrison, E. J. (1971). Performance and product evaluation. In Thorndike, E. L. (Ed.). *Educational Measurement (2nd. ed.)* (pp. 237-270). Washington,

DC: American Council on Education.

Fuchs, L. S., & Deno, S. L. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children, 19(8),* 1-16.

Fuchs, L. S., & Deno, S. L. (1991). Pardigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57(6),* 488-500.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation on achievement: A meta-analysis. *Exceptional Children, 53,* 199-208.

Fuchs, L., Fuchs, D., & Deno, S. (1982). Reliability and validity of curriculum-based informanl reading inventories. *Reading Research Quarterly, 18,* 6-25.

Gronlund, N. E. (1982). *Constructing achievement tests (3rd ed.).* Englewood Cliffs, NJ: Prentice Hall, Inc.

Hambleton, R. K. & Swaminathan, H. (1985) *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff.

Harney, W. & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys whats. and whithers. *Phi Delta Kappan, 70(9),* 683-687.

Jaeger, R. M. (1987). Two decades of revolution in educational measurement!? *Educational Measurement: Issues and Practices, 6,* 6-14.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365-377.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20(2),* 179-189.

Lovitt, T. (1967). Assessement of children with learning disabilities. *Exceptional Children, 34(4),* 233-239.

Marston, D. (1982). *The technical adequacy of direct, repeated measurement of academic skills in low achieving elementary students* (Unpublished Doctoral Dissertation). Minneapolis, MN: University of Minnesota.

Marston, D. & Deno, S. L. (1981). *Reliability of simple, direct measures of written expression (Research Report No. 50).* Minneapolis, MN: Institute for Research on Learning Disabilities, University of Minnesota.

Marston, D. & Magnusson, D. (1988). Curriculum-based measurement: District level implementation. In Graden, J. et al. (Eds.) *Alternative Educational Delivery Systems: Enhancing Instructional Options for All Students.* Washington, D.C.: NASP.

Messick, S. (1990). Test validity and the ethics of assessment. *American Psychologist, 35(11),* 1012-1027.

Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle. (1982). Frequency of measurement data and data utilization strategies as factors in standardized behavioral assessment of academic skill. *Journal of Behavior Assessment, 4(4),* 361-370.

Mitchell, S. R. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86,* 376-390.

Nathan, B. R & Cascio, W. F. (1986). Introduction: Technical & legal standards. In Berk,

R. A. (Ed.), *Performance assessment: methods and applications.* (pp. 1-50). Baltimore, MD: The Johns Hopkins University Press.

Neill, D. M. & Medina, N. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta kappan. 70(9),* 688-697.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18(9),* 3-7.

Nunnally, J. C. (1967). *Educational measurement and evaluation (2nd ed.).* New York: Free Press.

Popham, J. (1991). Interview on assessment issues with James Popham. *Educational Researcher, 20,* 24-27.

Priestley, M. (1982) *Performance assessment in education and training: Alternative techniques.* Englewood Cliffs, NJ: Educational Technology Publications.

Rogers, V. (1989). Assessing the curriculum experienced by children. *Phi Delta Kappan, 70(9),* 714-717.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal, 13,* 51-59.

Shavelson R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922-932.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46(7),* 4-9.

Shepard, L. A. (1991). Interview on assessment issues with Lori Shepard. *Educational Researcher, 20,* 21-23.

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children.* New York: Guilford.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1991). *Curriculum-based measurement reading fluency: A confirmatory factor analysis.* Unpublished manuscript. University of Oregon, Eugene.

Shinn, M. R., Tindal, G. A., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology. 3(1),* 69-85.

Smith, P. L . & Teeter, P. (1982). *The use of generalizability theory with behavioral observation.* Paper presented at the Annual Meeting of the AREA, 66th, New York, NY.

Stiggins, R. J. (1987, Fall). Design and development of performance assessments. *Educational Measurement: Issues and Practices.,* 33-42.

Stiggins, R. J. & Bridgeford, N. J. (1984). *Performance assessment for teacher development.* Northwest Regional Educational Lab, Portland, OR: Center for Performance Assessment.

Stiggins, R. J. & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement. 22(4),* 271-286.

Tindal, G., Marston, D. & Deno, S. (1983). *The reliability of direct and repeated measurement (Research Report No. 109).* Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6(1),* 1-12.

Webb, N. M., Rowley, G. L. & Shavelson, R. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21,* 81-90.

Weiss J. D., & Yoes, M E. (1988). Item response theory. In Hanbleton, R. K. & Zaal, Z. (Eds.), *New Development in Testing: Theory and Applications.* North-Holland Publishing Company.

White, O. R., & Haring, N. G.. (1980). *Exceptional Teaching (2nd ed.).* Columbus, OH: Merrill.

Wiggins, G. (1989). A true test: Toward more authetic and equitable assessment. *Phi Delta Kappan, 70(9),* 703-713.

Wolt, D., Bixby, J., Glenn, J., & Gardner, J. (1991). To use their minds well: Investigating new forms of student assessment. In Grant, G. (Ed.), *Review of Research in Education.*