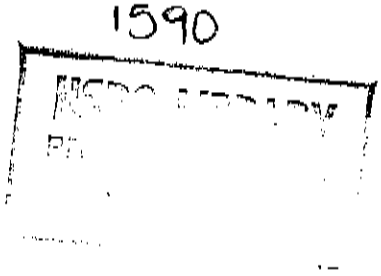
ASEESA CONFERENCE 2000

ASSESSMENT FOR COMPETENCY: TRANSFORMING POLICY INTO PRACTICE

HELD AT PORT ELIZABETH TECHNIKON

FROM SEPTEMBER 26th TO 29th 2000



PSYCHOMETRIC CHARACTERISTICS OF ALTERNATIVE ASSESSMENT TASKS

Mbithi wa Kivilu

Unit for Assessment Research and Technology
Human Sciences Research Council
Pretorious Street 134
Private Bag X41
Pretoria, 0001.

Abstract

Authentic assessment is a multifaceted process of testing learners by requiring them to show what they have learned in a context that is congruent with real-life experiences. Available methods in item analysis and test development rely heavily on classical test theory. The methods are designed for norm-referenced interpretation and may not be appropriate for determining psychometric characteristics in authentic assessment framework. Generalizability Theory (G-theory) offers study designs and efficient computational procedures that provide indices for both norm-referenced and criterion-referenced interpretation. In G-Theory factors that contribute to the measurement error can be identified and minimized, thus improving the validity of the measure. Data on an English Reading Test obtained from about 290 Grade 3 learners in in Gauteng- South Africa was used to demonstrate the application of G-theory in a authentic assessment framework. The English Reading Test had seven oral tasks and 10 reading tasks that were administered and scored by trained administrators. Variance components were determined for learners (object of measurement), types of task (oral versus reading), assessment task (or items), interactions and error terms. Index of dependability and reliability (Generalizability coefficient) indices were also determined. Suggestions are offered for the improvement of the English Reading Test for future use.

Psychometric characteristics of alternative assessment tasks.

Introduction

Demand for skills necessary for solving real life problems is putting pressure on education authorities to transform educational practices. Transformation in assessment practices has resulted in the emergence of a number of alternative assessment approaches. Recent education literature has promoted the value of alternative assessments also referred to as "performance assessment" or "authentic assessments" (Shepard, et al., 1996, Resnick & Resnick, 1992).

Several labels have been used to describe alternatives to standardized tests, with the most common being *direct assessment, authentic assessment, performance assessment* and the more generic *alternative assessment* which we shall use hereafter. Although these various descriptors reflect subtle distinctions in emphasis the several types of assessment all reflect two central commonalities. First, they are all viewed as alternatives to traditional multiple-choice, select-answer achievement tests. Second, they all refer to direct examination of learner performance on significant tasks relevant to life outside of school (Worthen *et al*, 1993).

Rise of Alternative assessment movement

Major trends in education have had a significant impact on the recent rise of alternative assessment movement:

- demands for higher standards and accountability in schools,
- work force for the technologically advanced world,
- use of test scores to make high-stakes decisions (promotion and graduation decisions),
- negative consequences of high-stakes testing programs. Pressures that accompanied high-stakes testing resulted in cheating where teachers coach learners on actual test items.
- Increasing criticisms of standardized tests. As the scores on such tests began to be used for increasingly crucial decisions, the test's limitations loomed larger.
- advances in cognitive and developmental psychological research (Worthen, *et al.*, 1993)

Alternative assessment is authentic if the process of testing learners require them to show what they have learned in a context that is congruent with real-life experiences. The demonstrations are usually performances linked to specific course exit standards (outcomes). The exit standards are the essential things in a given learning area that are required before a learner can move to the next grade, school or level of a training programme. Authentic assessment require clear and transparent articulation of criteria (standard) against which successful (or unsuccessful) performance is assessed. The criteria should specify the knowledge, understanding, performance(s), action(s) and roles that the learner needs to show in order to provide evidence that standards have been achieved. The criteria should also state the level of complexity and quality of the outcomes, standards and competence. The context of and conditions under which demonstration occurs should be indicated. Authentic assessments are multifaceted and measured over time, usually a year. In authentic assessment teachers often provide models, or benchmarks that can be viewed before hand.

In the debate over alternative assessment, the question being asked is not whether these tests are good assessment devices, but rather whether these tests can be used with the same

objectivity and understandability as multiple-choice tests. The testing situation is different in alternative assessment, it is usually longer and require more effort by both the teacher and the learner. The question of objective scoring is a real concern in authentic assessment. The main characteristics of an authentic assessment task (or activity) are:

- learner is required to perform a task using the skills and knowledge acquired during instruction,
- the standard of performance are stated in measurable terms
- it must be essential to learning,
- it can be measured over time
- has multiple forms for a response, and
- it should be closely tied to instruction

Unlike the traditional objective test items that have a unique answer, measurement of performance or scoring in authentic assessment involves some degree of subjectivity especially by raters. Error of measurement can also occur during the identification and definition of the expected standards or outcomes or level of performance. Thus for authentic assessment to provide quality results that are both reliable and dependable, administrators or raters should be trained to ensure consistency in scoring, measurement should be done on a number of occasions and interpretation of the results should be more of criterion- referenced than norm-referenced.

Technical quality and truthfulness of alternative assessment scores

There is currently little agreement about just what technical specifications and criteria should be used to judge the quality of alternative forms of assessment. Some assessment specialists would redefine or replace common conceptions of validity and reliability with alternative touchstones of acceptability (e.g. Wiggins, 1991), while others argue that alternative assessment will not be useful if its measures are not held to the same high standards of reliability and validity education has demanded of existing paper and pencil assessments (e.g. Cizek, 1991).

Thorny technical questions abound on a number of issues:

- Can one generalize satisfactorily from specific performance assessment tasks to the broader domain of achievement needs?
- Is performance task dependent or generalizable from task to task?
- How can assessment bias that has plagued traditional tests be kept from operating in alternative assessments that allow more subjectivity?

The crux of the issue is whether the alternative assessment movement will be able to evidence that its assessments are able to reflect accurately a learner's true ability in significant areas of behaviour of adult life. Whether called reliability, validity, or something else, some evidence that the technical quality of the assessment yields a truthful portrayal of learner abilities is essential.

Validity and Reliability Issues in Authentic Assessment

The issue of validity in assessment can be understood from two perspectives: (1) validity in measurement, precision of measurement, that is, the degree to which a measurement is free

from errors, and (2) validity in test use- interpretation of assessment scores for a specific use. Available methods in item analysis and test development rely heavily on classical test theory (CTT). Most of the methods in CTT are designed for norm-referenced interpretation and may not be appropriate for determining validity and reliability coefficients indices in authentic assessment framework

Classical test theory assumes strictly parallel measurement; that is, the means across items are assumed to be equal, as are the variances. Item effect is assumed to be zero. A consequence of the parallel-measurement assumption is that classical test theory is primarily a theory of individual differences, that is, it is usually concerned with the relative standing of individuals

The traditional classical test theory (CTT) conceptualizes observed scores as being made up of true score and an error component. The true score is a theoretical value that would be obtained if the test were administered to the same person infinite number of times under similar or equivalent conditions. The error component is supposed to be both random and systematic error and a conglomerate of other factors that are irrelevant to the measurement. In classical test theory the variations in individuals' observed test scores could be decomposed into only two components, namely, variation attributed to true differences among individuals, and variation attributable to systematic and random sources such as extraneous variables, interactions between the elements of measurement and the person's variables. Generalizability (G) theory has been used to dissect the sources of variation into all possible variance components of the measurement.

Generalizability (G) theory

Generalizability Theory (G-Theory) offers study designs and efficient computational procedures that provide indices for both norm-referenced and criterion-referenced interpretation. Generalizability (G) theory is a statistical theory about the dependability of behavioral measurement and uses the statistical technique of Analysis of Variance (ANOVA) to determine variance components. Dependability refers to the accuracy of generalizing from a person's observed score on a test or other measure (e.g., behavior observation, opinion survey) to the average score that person would have received under all possible conditions. A single score obtained on one occasion on a particular form of a test with a single administrator is not fully dependable, that is, it is unlikely to match that person's average score over all acceptable occasions, test forms, and administrators. The most serious sources of inconsistency or error or variability in measurement include, examinees, referred to as the object of measurement, test items, occasions, test forms, administrators (or raters), and interactions

In G-Theory, factors such as test items, occasions, test forms, and administrators are referred to as facets. Effects of the factors (facets) contributing to the measurement error can be identified and minimized, thus improving the validity of the measure. Generalizability theory offers an approach for assessing measurement consistency and the possibility of improving the reliability with which measurement are obtained while indicating the most efficient strategy for achieving desired measurement precision. The G-theory allows the decision maker to investigate the dependability (reliability in classical test theory) of scores for different kinds of interpretation, such as norm-referenced and criterion-referenced interpretations.

In G-theory a distinction is made between generalizability (G) studies and decision (D) studies. The purpose of a G-study is to anticipate the multiple uses of a measurement and to provide as much information as possible about the sources of variation in measurement.

G-studies attempts to identify and to incorporate into its design as many potential sources of variation as possible (universe of generalization). A D-study makes use of the information provided by the G-study to design the best possible application of the measurement for a particular purpose.

In planning a D-study the decision-maker (a) defines the universe of generalization;
(b) specifies the proposed interpretation of the measurement- relative (norm-reference) or absolute (criterion-referenced), the proposed interpretation defines measurement error and thereby identifies the sources of error of greatest concern; (c) uses the information from the G-study about the magnitude of the various sources of measurement error to evaluate the effectiveness of alternative designs for minimizing error and maximizing reliability. This evaluation is done in manner analogous to the Spearman-Brown prophecy formula in classical test theory. By increasing the number of conditions of a facet in a measurement the error contributed by that facet can be decreased, much as adding items to a test decreases error (and increases reliability) in classical test theory.

Apart from the task and rater facets, other facets need to be considered for performance assessments. These will include occasion, method and scoring rubrics and procedures. For comparability purposes scores should be stable from occasion to occasion. If method or mode of assessment (e.g. performance tasks, multiple-choice items, short-answer items) is incorporated as a facet in the universe of generalization, then generalizability theory often blurrs the distinction between reliability and validity (Brennan, 1992). Results invariant over mode of testing provide evidence of convergent validity and support the argument that different modes of testing provide exchangeable information. Some studies have indicated that this may not be so and that different modes of testing might provide different types of information about learner performance (Brennan, 1995).

If scores vary according to the particular scoring rubrics/procedures used, then the results cannot be meaningfully interpreted without a clear understanding of the specific rubric employed. Decision makers must understand not only what is being tested but also the standards and procedures used to assign scores.

Probably the most frequently discussed measurement issue when subjective scoring is involved is interrater reliability. Interater reliability coefficient considered in isolation can grossly exaggerate the dependability of scores on performance assessments. In the next section generalizability theory is discussed regarding its application to item analysis in authentic assessment

Generalizability theory provides two indices of reliability, one called generalizability coefficient is analogous to the classical test theory reliability coefficient and is used for norm-referenced interpretation (relative decisions). The second, is the index of dependability (Phi coefficient) which is used for criterion-referenced interpretation (or absolute decision). This reasoning for defining coefficients of generalizability and dependability extends to multifacet measurements (Shavelson & Webb, 1991)

Interpretation of Generalizability theory results (GENOVA Output)

Results from the analysis of English Proficiency Test (Oral & Reading) conducted with Grade 3 learners in Gauteng are presented for interpretation. The test comprised of 16 performance tasks for both oral and reading competency. Each of the 16 tasks was composed of varying number of items which were administered by teachers according to specified procedure. Factor analysis was used for confirmatory analysis by identifying items within a given domain. The test had construct validity because items within each domain marched with items in the related factor. Test administrators also referred to as rater were nested within learners. Thus the design of the study was Raters:persons x tasks ((r:p)xt)

The GENOVA program (Brennan, 1983) was used in determining the variance components and the coefficients of reliability and dependability. Table 1 shows the various sources of error in a two-facet measurement while Table 2 gives the variance components of a D-Study.

Table 1: Sources of Variability in a Two-Facet Measurement: Persons raters and cognitive tasks were all crossed: p x r x t

Source of variability	Types of Variation	Variance Notation
Persons (p)	Universe-score variance (object of measurement)	σ^2_p
Raters (r)	Constant effects for all persons due to stringency of raters	σ^2_r
Assessment task (t)	Constant effect for all persons due to their behavioral inconsistency from one task to another	σ^2_t
p x r	Inconsistencies of raters' evaluation of particular persons' performance	σ^2_{pr}
p x t	Inconsistencies from one task to another in particular persons' performance	σ^2_{pt}
r x t	Constant effect for all persons due to differences in raters' stringency from one cognitive task to another	σ^2_{rt}
p x r x t, *e*	Residual consisting of the unique combination of p, r, t; unmeasured facets that affect the measurement; and /or random events.	$\sigma^2_{prt,e}$

Interpretation of variance component in a Generalizability study

Both the coefficients of generalizability and dependability are high enough for decision making process, but can be improved by either increasing the number of items or learners. Practical considerations, for example, cost (money, time, and logistics) need to be assessed before a decision is made on the facet(s) to be modified.

Estimated Variance components from a generalizability study reflects the magnitude of error in generalizing from a persons score on a single item to his/her universe score (the persons average over all items in the universe). Estimated variance component depends on the scale of measurement, thus variance component are interpreted by their relative magnitude.

TABLE 2 Estimated Variance Components

Source of Variability	Estimated Variance Components	Percentage of Total Variance
[illegible]	438.86	66.7
[illegible]	16.37	2.5
[illegible]	124.33	18.9
[illegible]	67.38	10.3
[illegible]	10.26	1.6
[illegible]	0.82	
[illegible]	0.67	

The person variance component (66.7%) is an estimate of the variance across persons of person level mean scores where the mean is taken across all tasks and raters in the universe. This was the largest and expected given the diversity in the learner population in the district. The task component is the estimated variance of task mean scores where each mean is taken across all tasks and raters. This component accounts for 18.9% of the total variance. The amount of variance accounted for by this component suggests that the tasks differed somewhat in difficulty. The rater component is the variance of rater mean scores, where each mean is across persons and tasks. This component accounted for only 2.5% of the total variance. This implies that there is no significant variation among raters. Increasing the number of raters will not improve the reliability of the test in any significant way.

Of the interaction variance components, the largest is the one measuring person –by- task interaction which accounted for 10% of the total variance. This indicates that there are some differences in the ranking of person mean scores for each of the various tasks in the universe. The small percentage of variance accounted for by raters nested within persons indicated that they ranked the learners similarly.

Finally the last variance component is the residual variance that includes the triple-order interaction and all other unexpected sources of variation. This variance component accounted for only 1.6% of the total variance in mean scores due to interaction of persons, raters and tasks and/or other unsystematic or systematic sources of variance that were not measured

Conclusion

While authentic assessment seems to be suitable alternative to standardized achievement tests problems abound on providing evidence of both validity and reliability. The subjective nature of standards' setting and design of scoring rubrics coupled with variations in scoring exercise still present problems in the design of high quality authentic assessment. Although classical test theory is still useful in item analysis especially for objective tests, its inability to disentangle the various sources of variability in a given measurement makes inappropriate for determining psychometric characteristics of an authentic assessment task. Generalizability theory offers a solution but the complex nature of the computations involved makes it not accessible to most potential users.

References

Brennan, R. L. & Johnson E. G. (1995). Educational measurement: Issues and practices, Winter, 9-13.

Brennan, R. L. (1992). *Elements of generalizability theory (Rev. ed.).* Iowa City: American College Testing.

Cizek, G. J. (1991). Confusion effusion: A rejoinder to Wiggins. *Phi Delta Kappan,* 72(2) 150-153.

Resnick, L. B., & Resnick, D. P.. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford, & M. C. O'Connor (Eds). *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V. & Weston, T. J. (1996). Educational Measurement: Issues and practice, Fall 7-18.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools.* New York: Longman.. P 418-423

Wiggins, G. (1991). A response to Cizek. *Phi Delta Kappan,* 72(9), 700-703