1995

EDO-CG-95-25

ERIC Digest

# Fairness in Performance Assessment

*Tony C.M. Lam*

Performance assessment is the type of educational assessment in which judgments are made about student knowledge and skills based on observation of student behavior or inspection of student products (see the digest by Stiggins in this series). Education reformers have hailed policy that pushes performance assessment as manna (miraculous food) from above, feeding teachers and students "wandering in a desert of mediocrity" (Madaus, 1993, p.10). They claim that by replacing selection response tests such as multiple-choice tests with performance assessment, our schools will improve, and all ailments in student assessment, including the affliction of unfairness, will be cured. Unfortunately, although the pedagogical advantages of performance assessment in supporting instruction that focuses on higher order thinking skills are obvious, research has consistently indicated unresolved logistic and psychometric problems, especially with score generalizability (Linn, 1993). In addition, there is no evidence suggesting that assessment bias vanishes with performance assessment (Linn, Baker, & Dunbar, 1991).

### Bias & Fairness

Consonant with the unified conceptualization of validity (Messick, 1989), assessment bias is regarded as differential construct validity that is addressed by the question: To what extent is the assessment task measuring the same construct and hence has similar meaning for different populations? The presence of bias invalidates score inferences about target constructs because of irrelevant, non-target constructs that affect student performance differently across groups. These irrelevant constructs are related to characteristics, such as gender, ethnicity, race, linguistic background, socioeconomic status (SES), or handicapping conditions, that define the groups. For example, ability to read and understand written problems is a biasing factor in measuring mathematics skills because it is irrelevant to mathematics skills and it affects Limited English Proficient (LEP) and native English speaking students' performance differently on a mathematics test.

Assessment for its intended purpose is unfair if 1) students are not provided with equal opportunity to demonstrate what they know (e.g., some students were not adequately prepared to perform a type of assessment task) and thus the assessments are biased; 2) these biased assessments are used to judge student capabilities and needs; and 3) these distorted views of the students are used to make educational decisions that ultimately lead to limitations of educational opportunities for them. Despite a common definition of assessment fairness in reference to assessment bias, the approach and methods used to assure fairness are nevertheless determined by one's choice of either one of two antithetical views of fairness: equality and equity.

### Equality

The equality argument for fairness in assessment advocates assessing all students in a standardized manner using identical assessment method and content, and same administration, scoring, and interpretation procedures. With this approach to assuring fairness, if different groups of test takers differ on some irrelevant knowledge or skills that can affect assessment performance, bias will exist.

Traditional tests with selection response items have been criticized as unfair to minority students because these students typically perform less well on this type of test than majority students. However, no evidence is yet available to substantiate the claim that performance assessment can in fact diminish differential performance between groups (Linn et. al., 1991). Although the use of performance assessment can eliminate some sources of bias, such as testwisness in selecting answers that are associated with traditional tests, it fails to eliminate others, such as language proficiency, prior knowledge and experience, and it introduces new potential sources of bias: 1) ability to handle complex problems and tasks that demand higher order thinking skills (Baker & O'Neil, 1993); 2) metacognitive skills in conducting self-evaluation, monitoring thinking, and preparing and presenting work with respect to evaluation criteria; 3) culturally influenced processes in solving problems (Hambleton & Murphy, 1992); 4) culturally enriched authentic tasks; 5) low social skills and introverted personality; 6) added communication skills to present, discuss, argue, debate, and verbalize thoughts; 7) inadequate or undue assistance from parents, peers, and teachers; 8) lack of resources inside and outside of schools; 9) incompatibility in language and culture between assessors and students; and 10) subjectivity in rating and informal observations. (A strategy for reducing the influence of extraneous factors in rating that also supports integration of curricula is to employ multiple scales for different attributes embedded in the performance. For example, essays on social studies can be rated on subject matter knowledge, writing quality, and penmanship.)

With equality as the view of fairness, the strategy for reducing bias is to employ judgmental review and statistical analysis to detect and eliminate biased items or tasks. Recognizing the technical difficulties in statistical investigation of bias in performance assessment, Linn et. al. (1991) asserted that "greater reliance on judgmental reviews of performance tasks is inevitable" (p.18).

### Equity

Fair assessment that is equitable is tailored to the individual student's instruction context and special background such as prior knowledge, cultural experience, language proficiency, cognitive style, and interests. Individualization of assessment can be implemented at different levels in the assessment process, ranging from choice of assessment approach (e.g., a project instead of a test), content (e.g., selecting a topic to write an essay on, allowing translation), administration (e.g., flexible time, allowing a dictionary), scoring (e.g., differential weighting), and interpretation (e.g., using a sliding grading scale).

By assessing students using methods and administration procedures most appropriate to them, bias is minimized because construct-irrelevant factors that can inhibit student performance are taken into consideration in the assessment design. For example, in place of a paper-and-pencil word problem test in math to be administered to the class, a teacher could give the test orally to a LEP student, rephrasing the questions and using the student's native language if necessary. When assessment content is customized, congruence between assessment and instruction for all students is enhanced. And, by adjusting scoring and grading procedures individually based on student background and prior achievement, fairness is directly addressed.

Performance assessment, with its ability to provide students with rich, contextualized, and engaging tasks, can allow students to choose or design tasks or questions that are meaningful and interesting to them, can make adjustments based on student experiences and skills, and can test student individually "to insure that the student is fully examined" (Wiggins, 1989, p.708). These characteristics of performance assessment are indeed the major thrusts of equitable assessment. However, it is

the individualization strategy and *not* the performance task, that produces bias-free scores. If multiple versions of a multiple-choice test were written for students with varying learning experiences and backgrounds, and the test administered individually with opportunities for students to defend and explain their answers, similar results could be achieved. The persistent gap between majority and minority student performance on accountability tests, even after the introduction of performance-based sections, may be attributable partially to the fact that these tests are standardized.

The major difficulty in individualized performance assessment is assuring comparability of results. Student performance is not consistent across different contexts and topics in writing assessment, and across different experiments and assessment methods in science (see Miller & Legg, 1993). Attempts to develop tasks that are functionally equivalent have been scarce and unsuccessful. For example, it is difficult to construct comparable tasks of equal difficulty in writing assessment (Miller & Legg, 1993); methods of translating a test into another language and establishing the equivalence of scores are not well known and are used sporadically (Hambleton & Kanjee, 1993); and for constructed response exams that allow students to choose a subset of questions, it is not common in tryouts to have representative examinees answering all combinations of the questions (Wainer, Wang, & Thissen, 1994). Procedures for equating scores from disparate assessments are just as problematic. As noted by Linn & Baker (1993), "some desired types of linking for substantially different assessments are simply impossible" (p.2).

Other pitfalls in assuring equity in performance assessment through individualization strategies can also be noted. If students are delegated the responsibility of determining how they should be assessed, such as choosing an essay topic, picking out best work, or assigning points, individual differences in this metacognitive ability can become a source of bias. Furthermore, for any form of assessment, differential scoring and interpretation (such as the use of differential standards) encourage low expectations for the coddled students, and ultimately lessen their competitive edge when entering the workforce.

**Summary & Conclusion**

In dealing with the issue of fairness in performance assessment, we are confronted with some dilemmas. On the one hand, assuring equality in performance assessment through standardization enables comparisons of student performance and simplifies administration processes; however, it loses task meaningfulness and creates difficulty in avoiding bias. On the other hand, assuring equity effectively reduces bias and enables rich, meaningful assessment, but it introduces difficulty in administration and in comparing student performance, causes a potential side effect of poorly equiping students for the real world, and can be unfair to students with low awareness of their own abilities and quality of performance. Although standardized assessment is encouraged because it is a requirement for reliability, which is a necessary condition for validity, the hermeneutic approach to score interpretation supports contextualized and non-standardized assessment, and argues that validity can be achieved without reliability (Moss, 1994).

There is currently little research devoted to examining and promoting fairness in performance assessment. However, the urgency to build this knowledge base should not surpass the much needed research on, and efforts to develop, sound and practical performance assessments. When dealing with the issue of fairness in assessment, validity must be considered concurrently. How much better off are we with assessments that are equally invalid for all groups (fair but invalid) than assessments that are invalid for some groups (valid but unfair)?

**References**


Baker, E. L. & O'Neil H. F. (1993). Performance assessment and equity. In *Evaluating Education Reform: Assessment of Student Performance.* Washington, D.C.: Pelavin Associates.

Hambleton R. K. & Kanjee A. (1993). *Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods.* Paper presented at the annual American Educational Research Association Conference, Atlanta, Georgia.

Hambleton R.K. & Murphy E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5(1), 1-16.*

Linn, R. L. (1993). *Educational assessment: Expanded expectations and challenge.* CSE Technical Report 351, CA: CRESST.

Linn, R. L. & Baker, E. L. (1993). Comparing results from disparate assessments. *The CRESST Line,* CA.: National Center for Research on Evaluation, Standards, & Student Testing, 1-3.

Linn, R., E., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, G. (1993). A national testing system: Manna from above? An historical/technological perspective. *Educational Assessment, 1*(1), 9-26.

Messick S. (1989). Validity. In Linn R. (Ed.), *Educational Measurement,* New York: Macmillan Publishing Company, p.221-262.

Miller M. D. & Legg S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12*(2), 9-15.

Moss P. A. (1994). Can there be validity without reliability? *Educational Researcher, 32*(2), 5-12

Wainer H., Wang X., & Thissen D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? *Journal of Educational Measurement, 31*(3), 183-199.

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



*Tony C. M. Lam is Associate Professor, Faculty of Education, University of Toronto.*



**ERIC Digests are in the public domain and may be freely reproduced and disseminated. This publication was funded by the U.S. Department of Education, Of fice of Educational Research and Improvement, Contract No. RR93002004. Opinions expressed in this report do not necessarily reflect the positions of the U.S. Department of Education, OERI, or ERIC/CASS.**

For information on other ERIC/CASS products and services, please call toll-free (800) 414-9769 or (910) 334-4114 or fax (910) 334-4116 or write ERIC/CASS, School of Education, University of North Carolina at Greensboro, Greensboro, NC 27412.