DOCUMENT RESUME

ED 366 647 TM 021 061

AUTHOR	Taylor, Ronald D.
TITLE	Reassessing Performance Based Assessment.
PUB DATE	13 Nov 93
NOTE	18p.; Paper presented at the Annual Conference of the Missouri Unit of the Association of Teacher Educators (Osage Beach, MO, November 13, 1993).
PUB TYPE	Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Accountability; Cost Effectiveness; Cultural Awareness; *Educational Assessment; Educational Change; Elementary Secondary Education; Evaluation Methods; Literature Reviews; *Multiple Choice Tests; *Testing Problems; Test Reliability; *Test Use; Test Validity; Time Management
IDENTIFIERS	*Performance Based Evaluation; Reform Efforts; Testing Effects

ABSTRACT

A review of the recent literature has yielded a number of concerns about the validity, reliability, cost, efficiency, generalizability, utility, and cultural sensitivity of performance based assessments. The resulting conclusion was that continuing the performance based assessment initiative should be rethought. Suggested alternatives included understanding the mistaken rationale for abandoning multiple-choice testing, continuing the use of reformed multiple-choice tests, adopting a multiple measures approach to assessment, and recognizing the limits of testing in accountability and educational reform. (Contains 51 references.) (Author)

Reassessing Performance Based Assessment

Ronald D. Taylor

Columbia College

This paper includes information presented November 13, 1993, at the Annual Conference of the Missouri Unit of the Association of Teacher Educators, Osage Beach, MO.

Abstract

A review of the recent literature has yielded a number of concerns about the validity, reliability, cost, efficiency, generalizability, utility and cultural sensitivity of performance based assessments. The resulting conclusion was that continuing the performance based assessment initiative should be rethought. Suggested alternatives included understanding the mistaken rationale for abandoning multiple-choice testing, continuing the use of reformed multiple-choice tests, adopting a multiple measures approach to assessment, and recognizing the limits of testing in accountability and educational reform.

Reassessing Performance Based Assessment

In recent years, there has been a rush toward performance based assessment (PBA) as a reform to multiple-choice testing. Readings of the PBA literature over the past year, however, have led to the conclusion that an error has been made which should be remedied (Taylor, 1993). The purpose of this paper was to communicate why that conclusion was reached. Taking guidance from Bracey (1993a) and Lissitz and Schafer (1993), the objective was to demonstrate that there are sufficient concerns about validity, reliability, costs, efficiency, generalizability, utility and cultural sensitivity to warrant a reassessment of the PBA initiative.

No effort was made to balance the argument. A sufficient PBA literature has been amassed to obtain ample oppositional views. Similarly, no effort was made to distinguish between PBA and authentic testing. Both have been described as involving performance by the test taker and observation and measurement by the test giver. Thus, the terms were considered to be interchangeable.

The enormous costs (Rothman, 1993b) and the intent of states (Latus, 1993) and the federal government (U.S. Department of Education, 1993) to exclusively use PBA for accountability purposes make this an important issue. That those who have already used PBA have begun to retreat (Madaus & Kellaghan, 1993) heightens the importance for educators to more closely and critically examine continuance of this initiative.

Concerns About PBA

Validity

Gronlund and Linn (1990) have stated that imprecise instructional objectives may aversely influence the validity of tests designed to measure intended outcomes. The lack of precision in outcomes guiding PBA have been widely criticized (Harp, 1993a, 1993b; Kirst, 1991; Pipho, 1992; Shanker, 1993; State Journal, 1993). It is not surprising, therefore, that PBA validity evidence has been described variously as marginal (Lissitz & Schafer, 1993), primarily assumptive (Linn, Baker & Dunbar, 1991), little found (Baker, O'Neill & Linn, 1991), little agreement about (Worthen, 1993), lacking as a fundamental property (Brooks & Parker, 1993), and threatened by costs (Worthen, 1993), as well as teacher coaching (Madaus & Kellaghan, 1993). Educational leaders in Maryland, referring to their recent PBA experience, suggested that results were so flawed they should be thrown out (Flax, 1992). Finally, Brown (1993) has reported that PBA has not yet demonstrated validity equal to that found in standardized, multiple-choice tests.

Reliability

PBA in Britain, Wales and Scotland was initiated in 1991 (Madaus & Kellaghan, 1993), and the instruments used are known as the British Standardized Assessment Tasks (BSATs). Schmidt (1993) reported, as of yet, BSATs have failed to demonstrate reliability. In this country, Shepard (Cohen, 1993) has admitted that certain PBAs have not demonstrated sufficient reliability for external accountability purposes. Similarly,

Rothman (1991a) has reported that the reliability of PBA in Vermont was so low it seriously limited the usability of results.

Just as imprecise instructional outcomes may aversely affect test reliability, they also have been described as threatening to validity (Gronlund & Linn, 1990). Outcomes accompaning PBAs have been described as: (a) vague and fluffy (Shanker, 1993); (b) vague and ill defined (Pipho, 1992); (c) less specific and clear (Kirst, 1991); (d) soft (State Journal, 1993); (e) subjective (Harp, 1993a, 1993b); and (f) threatened by costs (Popham, 1993a; Worthen, 1993).

Others have described PBA reliability evidence as little found (Baker et al., 1991), little agreement about (Worthen, 1993), lacking as a fundamental property (Brooks & Pakes, 1993), yet to be demonstrated (Brown, 1993), and marginal (Lissitz & Schafer, 1993). Lastly, and more kindly, Barth (1993) reported PBS reliability data as inconclusive.

Costs and Efficiency

Regarding costs, Popham (1993a) candidly stated about PBA, "In short, it takes money -- a lot of it" (p. 472). Clearly, there has emerged a consensus about the costliness of PBA (Diegmueller, 1993; Harp, 1993b; Kirst, 1993; Madaus & Kellagham, 1993; Worthen, 1993). Rothman (1993b), having analyzed national data, reported the average cost for administering a multiple-choice test to be $16; but, for a PBA, the average cost was estimated at $33 per test. For the nation, PBA development costs have been estimated to be $100 million and administration costs to be $330 million per

year. Others (Madaus & Kellaghan, 1993) have estimated less annual cost than Rothman (1993), $115.5 to $175 million, but others have estimated more, $750 million to $1.5 billion (Madaus & Kellaghan, 1993). Whomever is more correct, the costs may be termed staggering.

It should be noted also that as a strategy to offset the PBA costs, Popham (1993a) has suggested the use of true matrix sampling. This technique involves item sampling and student sampling. Using this approach, a "limited number of tasks" (p. 473) would be given to the "smallest number of students" (p. 473). Thus, fewer students would be tested over less content, but money would be saved. Adopting such a strategy would require caution. Any sampling errors introduced would threaten further validity and reliability. Also, such an approach rests upon the assumption that teaching would be affected as much by what potentially could be tested rather than what actually has been tested. History would suggest this a risky assumption. Lohman (1993) noted that untested outcomes tend to go untaught.

Efficiency of PBA also has been characterized as enormously impaired, especially with respect to time commitments (Brown, 1993; Cohen, 1993; Madaus & Kellaghan, 1993). For example, Madaus and Kellaghan (1993) reported that, to administer the BSATs to a class of 36 students, an estimated half school term would be necessary. In addition to time considerations, the BSATs were reported to be enormously disruptive to routines, organization and personnel (Madaus & Kellaghan, 1993). Indeed, Schmidt (1993) reported the experience so upsetting that all but a

handful of some 4,400 teachers had vowed to boycott future BSATs.

Generalizability

Greeno (1989) has reported that performance tasks are specific rather than generalizable. Shavelson, Baxter and Pine (1992) have agreed and have labeled science PBA tasks poorly generalizable. Worthen (1993) has concluded that four important questions about generalizability have yet to be answered:

1. Can generalizations be made from specific performance tasks to the broader domain of achievement?
2. What does completion of a specific project reveal about knowledge otherwise in that particular content area?
3. What types, and in what amount, of performance tasks are required before appropriate generalizations can be made?
4. Are performance tasks sufficiently isomorphic and how can we know with certainty?

Linn et al. (1991) were correct -- PBA will not lessen our concerns about generalizability.

Utility

How useful will PBA results be? To what uses may PBA results be applied? PBA data may not yield the answers desired. Madaus and Kellaghan (1993) have reported that British educators believed the BSATs added little to nothing to what was already known about students. Diegmueller (1993) reported that Pennsylvania

legislators were concerned that outcomes based education, assessed by PBA, would result in lowered expectations and ineffectual instruction. Eakman (1993) has recently discussed other concerns about the Pennsylvania Educational Quality Assessment published by the Educational Testing Service in 1984.

Cuban (1993b) has warned that minority, poor children receive few benefits from any systematic, educational reform in general. For PBA specifically, Linn et al. (1991) reported it would be wrong to assume that PBA will result in more conducive learning activities for any students, or that PBA will result in improved problem solving and reasoning skills.

Bracey (1993b) has discussed elsewhere the issue of educational corruption. Could one assume PBA will be less suspectable to corruption? Linn et al. (1991) reported the answer would be no -- that it would be wrong to assume PBA's immunity to corruption. Could we be more certain about PBA's assessments of students' knowledge and skills? No, Airasian (1993) concluded PBA will not permit us to lessen assessment pitfalls and uncertainties. Finally, could PBA provide superior insight regarding student performance than that provided through standardized achievement tests? Results reported by Madaus and Kellaghan (1993) suggest not. Shavelson et al. (1992) also have suggested not, saying it would be erroneous to assume that educators will know more using PBA than that provided by standardized achievement tests.

Cultural Sensitivity

Fitzgerald (1993) has reported that new test formats will not remedy gender and racial biases. Regarding PBA, Linn et al. (1991) have reported it would be wrong to assume a lack of bias against racial/ethnic minorities. Instead, it has been argued that PBA makes more difficult the task of developing culturally contextualized assessments (Linn et al., 1991). Brown (1993) has concluded that subjectivity in scoring PBAs will increase the burden of bias due to culture, sex, language and content. In fact, Dunbar, Koretz and Hoover (1991) have concluded flatly that PBAs widen the gap between the majority group and various minority groups. Finally, Brown (1993) has observed that minorities may be more discriminated against by PBAs than by multiple-choice tests.

Discussion

The recent literature has suggested sufficient concerns for a reassessment of the PBA initiative. Validity and reliability evidence for PBAs appears questionable. Clearly, PBA will be substantively more costly and less efficient and utilitarian. The important questions about generalizability have not been answered, and there is evidence PBA will exacerbate, rather than lessen, cultural sensitivity problems.

Educators and educational reformers should rethink PBA. Why have we elected to use a prehistoric testing technique and cast it as our future psychometric salvation? The fundamental reason why PBAs have been embraced was based on the mistaken conclusion that the multiple-choice testing format had corrupted what was

taught students, and, consequently, limited students' high category abilities and skills.

What led to this mistaken conclusion? Since the National Commission on Excellence in Education Report (1983), multiple-choice tests have been employed extensively to test for minimum competencies (low category abilities and skills) in basic skill areas (Airasian, 1993; Brown, 1993; Cuban, 1993; Jett & Schafer, 1993; Madaus & Kellaghan, 1993). Teachers, therefore, have concentrated on training those abilities and skills (Airasian, 1993; Brown, 1993; Cuban, 1993a; Jett & Schafer, 1993; Rotberg, 1993), an expected outcome (Lohman, 1993; Popham, 1993a, 1993b). Thus, it has been the content of multiple-choice tests (Brown, 1993; Rotberg, 1993) and their intended purpose (Brown, 1993) that have resulted in common complaints about testing influences on teaching. It was not the fault of the test format (Brown, 1993).

Multiple-choice tests can assess both low and high category abilities and skills (Anastasi, 1988; Gronlund & Linn, 1990). The only limitations to this format are the skills of test writers and test users. We only need to write multiple-choice tests that sample what we want teachers to teach and students to learn, the entire range of categorical abilities and skills. We then can expect to see teachers respond in the classroom (Airasian, 1993; Madaus & Kellagham, 1993; Popham, 1993a, 1993b), and we will have more valid, reliable, cost and time efficient, and useable measures. Indeed, the lack of these characteristics led to the abandonment of performance testing in the first place (Madaus & Kellaghan, 1993; Popham, 1993b). Finally, while multiple-choice tests have not overcome biases due to culture, sex, language and content

(Brown, 1993), there is increasing evidence that gaps are being narrowed or stabilized (Shea, 1993a, 1993b). And, we have known for some time that selected-response examinations correlate quite well with constructed-response examinations (Popham, 1993b), permitting us some confidence in generalization of their results.

The data presented herein, however, should not be taken as advocacy for multiple-choice measures exclusively. Kean (1993) and Stiggins (1993) have cogently argued for the adoption of a multiple measure approach. There is a need to fit the measure to the task. When measuring what students should know, paper-and-pencil testing may be employed. When measuring what students should be able to do, performance measures may be employed. On its face, this seems a most reasonable and logical approach.

Lastly, the limited role of testing of all types in providing accountability must be recognized, as must be recognized the limited effect of testing on educational reform. Hansen (1993) has said it best:

> Using accountability . . . as the primary means to achieve reform is somewhat akin to using a tape measure as the treatment to reduce one's waist size. Frequent, accurate monitoring, using reliable and valid assessment tools, is necessary to measure progress, but it will not, in and of itself, cause desired change. (pp. 19-20)

References

Airasian, P.W. (1993). Policy-driven assessment or assessment-drive policy. Measurement and Evaluation in Counseling and Development, 26, 22-30.

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.

Baker, E.L., O'Neill, H.F. Jr., & Linn, R.L. (1991, August). Policy and validity prospects for performance based assessments. Paper presented at the 99th Annual Convention of the American Psychological Association, San Francisco.

Barth, P. (1993, October 2[illegible]). The education holdup: Peg Luksik and the right want to back schools into the future. Education Week, pp. 39, 43.

Bracey, G.W. (1993a). Assessing the new assessments. Principal, 72, 34-36.

Bracey, G.W. (1993b). The third Bracey report on the condition of public education. Phi Delta Kappan, pp. 104-112, 114-117.

Brown, D.C. (1993). America 2000 and policy implication for the country. Measurement and Evaluation in Counseling and Development, 26, 48-53.

Brookes, T.E., & Pakes, S.J. (1993, October). Policy, national testing, and the Psychological Corporation. Measurement and Evaluation in Counseling and Development, 26, 54-58.

Cohen, D.L. (1993, January 20). Assessment alternatives for younger students seen honing teacher's skills, observations. Education Week, pp. 6-7.

Cuban, L. (1993a, July 14). A national curriculum and tests: Charting the direct and indirect consequences. Education Week, p. 25.

Cuban, L. (1993b, October). The lure of curricular reform and its pitiful history. Phi Delta Kappan, pp. 182-185.

Diegmueller, K. (1993, February 17). PA house votes to nullify board's learner outcomes rule. Education Week, p. 19.

Dunbar, S. , Koretz, D., & Hoover, H.D. (1991). Quality control in the development and use of performance assessment. Applied Measurement in Education, 4, 289-303.

Eakman, B.K. (1993, October 20). It's about mental health, stupid. Education Week, pp. 40, 43.

Fitzgerald, M. (1993, October 12). Admissions tests hold sway despite questions. Columbia Daily Tribune, p. 8A.

Flax, E. (1992, June 3). Maryland should invalidate the results of new testing program, teachers say. Education Week, p. 20.

Greeno, J.G. (1989). A perspective on thinking. American Psychologist, 44, 134-141.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hansen, J.B. (1993). Is educational reform through mandated accountability an oxymoron? Measurement and Evaluation in Counseling and Development, 26, 11-21.

Harp, L. (1993a, September 15). Groups fighting outcomes plan in Washington at odds. Education Week, p. 22.

Harp, L. (1993b), September 22). PA parent becomes mother of "outcomes" revolt. Education Week, pp. 1, 19-21.

Jett, D.L., & Schafer, W.D. (1993). Ready or not, teachers K-12 move to center stage in the assessment arena: Implications for state education policymakers. Measurement and Evaluation in Counseling and Development, 26, 69-80.

Kean, M.H. (1993, October 6). Getting it right: Authentic assessments and the true multiple measures approach. Education Week, pp. 27, 29.

Kirst, M. (1991a). Interview on assessment issues with Lorrie Shephard. Educational Researcher, 20, 21-23, 27.

Kirst, M. (1991b). Interview on assessment issues with James Popham. Educational Researcher, 20, 24-27.

Latus, J. (1993, Fall). Outstanding schools, Missouri style. School and Community, pp. 8-13.

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

Lissitz, R.W., & Schafer, W.D. (1993). Policy-driven assessment: An old phenomenon with new wrinkles. Measurement and Evaluation in Counseling and Development, 26, 3-5.

Lohman, D.F. (1993). Teaching and testing to develop fluid abilities. Educational Researcher, 22, 12-23.

Madaus, G.F., & Kellaghan, T. (1993a, February). The British experience with authentic testing. Phi Delta Kappan, pp. 458-459, 462-463, 466-469.

Madaus, G.F., & Kellaghan, T. (1993b). Testing as a mechanism of public policy: A brief history and description. Measurement and Evaluation in Counseling and Development, 26, 6-10.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U.S. Department of Education.

Noll, V.H. (1955). Requirements in educational measurement for prospective teachers. School and Society, 82, 88-90.

Pipho, C. (1992, May). Outcomes or edubabble. Phi Delta Kappan, pp. 662-663.

Pitsch, M. (1993, July 14). House approves $28.6 billion budget for education. Education Week, p. 22.

Popham, W.J. (1993a, February). Circumventing the high costs of authentic assessment. Phi Delta Kappan, pp. 470-473.

Popham, W.J. (1993b). Measurement-driven instruction as a "quick-fix" reform strategy. Measurement and Evaluation in Counseling and Development, 26, 31-34.

Rotberg, I.C. (1993, September 29). Chapter I testing: Its no field of dreams. Education Week, p. 40.

Rothman, R. (1993a, January 13). VT revises assessment in wake of study finding problems. Education Week, p. 31.

Rothman, R. (1993b, February 3). Testing places only a "modest" burden on student, G.A.O. report concludes. Education Week, p. 25.

Rothman, R. (1993c, February 17). Putting it to the test. Education Week, pp. 1, 22.

Schmidt, W.E. (1993, August 1). Britain flunks a test of its national curriculum. Education Life, pp. 17, 19.

Shanker, A. (1993, September 28). Outcome-based plan too fluffy. Columbia Daily Tribute, p. 8A.

Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political, rhetoric and measurement reality. Educational Researcher, 21, 22-27.

Shea, C. (1993a, September 1). Average SAT scores rose in 1993 for the second year in a row. The Chronicle of Higher Education, p. A46.

Shea, C. (1993b, September 22). Average ACT score for 43 graduates increases by 0.1. The Chronicle of Higher Education, p. A34.

State Journal. Soft goals. (1993, September 29). Education Week, p. 18.

Stiggins, R.J. (1993). Two disciplines of educational assessment. Measurement and Evaluation in Counseling and Development, 26, 93-104.

Taylor, R.D. (1993, November 13). Back to the future in assessment: Problems with performance based assessment. Paper presented at the Annual Conference of the Missouri Unit of the Association of Teacher Educators, Osage Beach, MO.

U.S. Department of Education (1993, October 20). Improving America's schools act of 1993: The reauthorization of the elementary and secondary education act and amendments to other acts. Education Week, pp. 19-34.

Worthen, B.R. (1993, February). Critical issues that will determine the future of alternative assessment. Phi Delta Kappan, pp. 444-454.