DOCUMENT RESUME

ED 286 904 TM 870 535

AUTHOR McLean, Les
TITLE Emerging with Honour from a Dilemma Inherent in the Validation of Educational Achievement Measures.
PUB DATE Apr 87
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Behavioral Objectives; Educational Assessment; Elementary Secondary Education; Evaluation Methods; *Evaluation Problems; *Instructional Effectiveness; Instructional Improvement; Mathematics Tests; *Measurement Objectives; Native Language Instruction; Portfolios (Background Materials); Second Language Learning; Student Evaluation; *Validity
IDENTIFIERS Relevance (Evaluation); *Second International Mathematics Study

ABSTRACT

A dilemma arises in the attempt to establish procedures for valid assessment of academic learning. Measures of learning with high pedagogical validity often have poor psychometric properties. Conversely, tests that are well-constructed by psychometric standards may succeed in sorting and ordering students or schools, but have little pedagogical relevance. There are several approaches for dealing with apparent trade-off. One way is to revise notions of achievement, moving away from test-based criteria toward performance-based criteria. Experience in native- and second-language learning is cited as instructive, because of the multidimensional emphasis on the goals of communication in realistic settings. Valid assessment depends on situation and mode of communication. A systematic, cumulative record of performance is essential to this approach. The writing folder, a dossier kept by the student of diverse samples of his or her writing, is a good example of such a record. Another approach connects the monitoring function of assessment to the learning function at the classroom level. It uses comprehensive item pools administered by item sampling as part of a survey of teaching and learning specifically designed to reflect school learning and suggest improvements. As an example, the Second International Mathematics Study is discussed. (LPG)

Emerging with Honour from a Dilemma Inherent in the Validation of Educational Achievement Measures

Les McLean
Ontario Institute for Studies in Education

Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 20-24, 1987.

Emerging with Honour from a Dilemma Inherent in the Validation of Educational Achievement Measures[1]

Les McLean
Ontario Institute for Studies in Education

Educational achievement is a slippery concept, with the result that the meaning of educational achievement measures is always difficult to estsblish. It is the classic validity problem--justifying interpretation of mental measurements. Anastasi (1986) expressed the psychologist's point of view as follows,

> In educational contexts, the most characteristic tests are the *so called* achievement tests, whose purpose is to assess the effects of academic learning and the individual's readiness for further learning of a similar nature (p. 7, emphasis added).

Carroll (1987) added the other dimension most often associated with achievement measurement,

> The purpose of national assessments of reading is to give the education community and the general public a total view of the state of reading literacy in the nation, *with enough detail to enable both groups to draw inferences about what steps might be taken to improve that state* (p. 426, emphasis added).

The focus in this paper is on the assessment of academic learning, school learning in particular, concentrating on assessments that suggest ways to improve that learning. Valid assessment of academic learning separates academic from other learning and reflects schooling effects much more than, say, conditions in the home or community. Achieving validity of this type does not, however, guarantee that the assessment will suggest ways to improve learning. Ideal assessments reflect school learning *and* suggest improvements--in other words, they have high validity in both senses. They have pedagogical relevance (McLean, 1986a), so we will refer to this as *pedagogical validity*. Well-constructed tests can yield scores that have validity for sorting and ordering students (hence for marking) but do not have pedagogical validity.

Lest we forget, sorting and ordering is precisely what is desired for scholarship examinations, university entrance and like prize competitions. Moreover, if test scores reflect the most valued outcomes of secondary schooling (a very large *if*), then such tests are appropriate for diploma examinations, such as the ones that account for 50 percent of the graduation mark in 5 of the 10 Canadian provinces. In other words, where summative evaluation is the goal rather than explanation and improvement, test scores measuring composite traits can be efficient tools. What has become clear, however, is that their usefulness stops with sorting and ordering. They can also be used to sort

---

[1] Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 20-24, 1987.

classes and schools, but their valid use for that purpose is a research effort of considerable complexity, requiring measures of school and class characteristics (Aitkin & Longford, 1986).

The Dilemma

Trying for pedagogical validity, however, confronts us with a dilemma. As we will see, assessments connected to teaching and learning, assessments that closely reflect school effects and yield suggestions for improvement, have low quality ratings on psychometric criteria. Measures with good ratings by psychometric criteria have low validity, and the higher the ratings, the lower the pedagogical validity. Do we accept lower ratings or lower validity? Fortunately, there is an honourable way out. Finding that way means going back to first principles, principles of teaching and learning that is, and leads us to question the value of some widely-accepted notions. For example, achievement constructs (Haertel, 1985) would appear to have limited utility.

Finding A Way Out

Setting out to assess academic learning begins with a definition of academic learning--and with a definition of achievement. Following the lead of the psychologists of the thirties, achievement has been defined in terms of test scores. Criterion-referenced testing was supposed to bring us closer to the substance of education, but domain descriptions and the like have been quickly converted to scores. "After defining the construct, the next step is to design and construct a test" (Haertel, 1985, p. 39). According to this view, achievement is a trait to be given operational definition by a set of items. Defining achievement in this way brings with it the elaborate apparatus of testing and test theory, reaching its apogee in item response theory.

This is not the only way to construe achievement, of course, as artists, musicians and athletic coaches are fond of reminding us. In these domains, achievement is, or should be, defined primarily in terms of performance. Until recently, they were the few nagging exceptions to the rule. Now, a break with trait measurement is happening in a core academic area, languages--mother tongue and second languages. The break came first in second language classrooms when parents and others demanded that students learn to communicate in realistic settings, especially to speak and understand languages other than their mother tongue.[2] It is coming in mother tongue classes because of dissatisfaction with student progress. Language structure has not been abandoned, but it has been made the servant of communication rather than the master.

---

[2]The term, *foreign language*, is not appropriate for French in Canada, Switzerland or Belgium, or for other languages elsewhere, hence the term, *second languages*. Since it may be the third or fourth for some people, a better term is simply, *other languages*.

Communicative, or functional, theories now dominate both the teaching and testing of language. According to these theories, language achievement is a complex, multidimensional, task dependent, situation dependent, person dependent performance (Gorman, 1986, Thornton, 1986). Language achievement, therefore, is located in a region of some multidimensional space far away from the black hole of trait theory.[3] David Olson expressed it this way,

> There used to be a thing like verbal ability that explained things. It's gone. At least it's gone for me. All cognitive science is based on the notion that you have set procedures that you use for dealing with domains. Some of those procedures are applicable across domains and so on. And the task for psychologists, as I think for educators and others, is to find out just what that competence is made up of and what are the conditions under which you can help people sort out the major dimensions or considerations in that form of competence. There is still talk about spatial ability and verbal ability and so on, but that's a level of description that has very little explanatory value. Explanatory value comes from actually figuring out how they solve this task, or how they sort out what next to put down on their text if they're writing something. (Olson, 1986, p. 177).

Olson might well have added that *reading ability* and the other *abilities* are gone for everyone who wants to know how to make things better, not just sort and order students on an abstract scale (McLean & Goldstein, 1987).

The exciting thing about functional language theories is that they contain a definition of achievement and suggest how that achievement should be attained--and measured. Meaning depends on context, and the *way* one communicates depends on the situation, the person(s) with whom one is communicating and the mode of communication. Thus, the language assessor has to specify situations in some detail and cover several modes in order to have valid assessment. There is no substitute for a systematic, cumulative record of performance.

A good example of such a record is the *writing folder*--a dossier kept by the student, into which go diverse samples of the student's writing, some marked by the teacher, some by the student and some by other students. The folder travels with the student from grade to grade. Its use is now mandated by the Ontario Ministry of Education in all elementary and secondary grades. Such a record is easily connected to learning and teaching. It has pedagogical relevance and, because it is faithful to the theory, it has pedagogical validity. Alas, it is messy, idiosyncratic and unstable--in other words, psychometrically hopeless. What some of us have long suspected is that classical and modern test theory is pedagogically hopeless. The assumptions required to make the theory usable disconnect the test scores from learning and reduce pedagogical validity almost to zero. Language theorists and teachers have shown us a possible way out.

[3] Black holes are postulated objects whose mass is so intensely concentrated that even light cannot escape. Most properties of the black hole are not observable.

And what of mathematics and science? We in the Western world can be old in our thinking, because all is certainly not well with schooling as it is now. One reads of the crisis in science education (Duschl, 1985), and "science is rarely taught adequately (if at all) in elementary schools across the country" (Science Council, 1984, p. 10). A pessimistic report on US results in the Second International Mathematics Study was entitled, *The Underachieving Curriculum* (McKnight et al., 1987). The lessons being learned in language classes deserve consideration in science and mathematics classes as well (McLean, 1986b).

Perils Along the Way

Dangers lurk along the functional path, however, one of the greatest being the threat to accountability. By its very nature, it is difficult to summarize a systematic, cumulative record of performance in any simple way. It is quite against the spirit of both theory and pedagogy to assign a mark to every entry and then average the marks. Moreover, such records are not easily compared, since they will contain different amounts of work, of different types--another reason not to calculate averages that beg to be compared. The path of individualization was never smooth (Suppes, 1964). Parents who can see students' writing folders, however, and who receive a brief evaluation from the teacher (perhaps backed up by the principal), are usually satisfied that justice is being done. The problem really lies with the hardy perennials--standards and fairness.

What should a record contain by the end of Grade 6 (or 9, or 12)? How much is enough? What constitutes superior work? How can we be assured that teachers are asking enough of the students and, even more important, how can we ensure that students are treated fairly? The best answer to all these questions has never changed--depend on good performance from well-prepared, motivated teachers. The rejoinder to that has not changed either--even good teachers need to know where they stand, and not all teachers are good. We will argue that finding out where one stands should be a process separated from the classroom (but not from the curriculum), and that within the classroom, there are fewer threats to the validity of achievement as performance than achievement as trait.

Emerging with Honour

Pedagogical validity is worth having. It may be hyperbole to call it the *soul* of assessment, but that name conveys the right spirit. Gaining accountability with low validity tests (however high their psychometric quality) profits us not *if* we lose our soul, but accountability *is* indispensable. If pedagogical validity is the soul, accountability is the body, the visible manifestation of assessment. We can have a healthy body and preserve the soul, but to do so we have to separate assessment for the purpose of teaching and learning from assessment for the purpose of monitoring. We can keep the

monitoring function connected to learning at the classroom level-- a neat trick, but we know how to do it. It is done with comprehensive item pools administered by item sampling as part of a survey of teaching and learning. The survey must permit aggregation to the operational level, usually the classroom but sometimes the school, and the survey must include measures of opportunity to learn. We know how to do this, because we did a good approximation of it in the Second International Mathematics Study (SIMS) and because SIMS showed us how to do it better next time. Monitoring can also have pedagogical validity.

Recall the requirements for pedagogical validity--reflect *school* learning and suggest improvements. Reflecting school learning is accomplished in two ways, (a) by the design of the item pool and (b) by the design of the survey. The pool can attain comprehensive coverage of the curriculum, because no student has to answer more than a fraction of it (15 to 20 percent) and because we have the technology and the resources to include a variety of items other than multiple choice. At the classroom level, the responses can be aggregated over items to subsets still connected to learning. The resulting achievement data can be linked to opportunity to learn and other class and school characteristics using multilevel statistical models whose complexity at least approximates the complexity of schooling. Because the results are still connected to learning, suggestions for improvement abound.

An important lesson about the design of the survey was learned from SIMS. If we really want to reflect school learning and suggest improvements, we should measure achievement at both the beginning and the end of the school year, keeping both linked to their classrooms. The payoff is worth the extra cost, because the pretest-posttest design allows us to look at achievement within the year, taking account of prior experience. This *look* is best done at the item level.

Many will be shaking their heads, having concluded that such an ambitious monitoring scheme is hopelessly impractical as an ongoing accountability tool. The start-up costs are indeed significant, but the operation can then be continued at no greater cost than now spent for standardized tests--provided you can keep the organization that did the survey intact. Nine Ontario boards of education (school districts) have now repeated much of the SIMS methodology in their own jurisdictions, some of them twice, at a cost of about $1.50 (US) per pupil, with the service being provided by the Educational Evaluation Centre of the Ontario Institute for Studies in Education, the organization that mounted SIMS in Ontario.[4] Officials in these boards will attest to the accountability value of the surveys, and to

[4]Ontario and British Columbia participated in SIMS as *countries*, as did Scotland, French Belgium and Flemish Belgium.

their pedagogical validity as well. Surveys cannot serve as ongoing evaluation tools because they do not yield individual student assessments, but with valid monitoring in place, that job can be left with confidence to the teachers and the schools.

References

Aitken, M., & Longford, N. (1986). Statistical modelling in school effectiveness studies . *Journal of the Royal Statistical Society, 149*, 1-43. With discussion.

Anastasi, Anne (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Carroll, John B. (1987, Feb.). The national assessments in reading: Are we misreading the findings? *Phi Delta Kappan, 68*(6).

Duschl, Richard A. (1985). Science education and philosophy of science--twenty-five years of mutually exclusive development. *School Science and Mathematics, 85*(7), 541-555.

Gorman, Tom. (1986). *The framework for the assessment of language.* Windsor, Berkshire: NFER-NELSON.

Haertel, Edward (1985, Spring). Construct validity and criterion-referenced testing. *Review of Educational Research, 55*(1), 23-46.

McKnight, Curtis C., Crosswhite, F. Joe, Dossey, John A., Kifer, Edward, Swafford, Jane O., Travers, Kenneth J., & Cooney, Thomas J. (1987, January). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective.* Champaign, IL: Stipes Publishing Company.

McLean, Les. (1986, June 25-27). The search for pedagogical relevance in large-scale assessment. Paper presented at the Invitational Seminar, International Approaches to the Assessment of Student Performance meeting of the National Foundation for Educational Research and University of London Institute of Education.

McLean, Les. (1986, September). Function and structure--the basis for real reform in curriculum. Paper presented at the Policy Dimensions of Curriculum Proposals meeting of the Canadian Association for Curriculum Studies, St. Andrews, N.B. With discussion.

McLean, Les, & Goldstein, Harvey (1987). *The U.S. national assessments in reading: Reading too much into the findings.* Unpublished manuscript, OISE.

Olson, David R. (1986). Mining the human sciences. *Interchange, 17*(2), 159-177. With discussion.

Science Council of Canada. (1984). *Report 36. Science for every student: Educating Canadians for tomorrow's world.* Ottawa: Minister of Supply and Services, Canada.

Suppes, Patrick (1964, March). Modern learning theory and the elementary-school curriculum. *American Educational Research Journal, 1*(2), 79-94.

Thornton, Geoffrey. (1986). *APU language testing 1979-1983: An independent appraisal of the findings.* London: Department of Education and Science.