

Running head: MINISTRY OF EDUCATION SYMPOSIUM 2002

Item types and validity/reliability in large-scale assessment: *The Emperor isn't wearing any
clothes*

by W.A. (Bill) Angus, M.A.

Student, UBC faculty of Education (MERM)

Requests for reprints should be sent to: W.A. (Bill) Angus, M.D. Angus & Associates Ltd., 2639
Kingsway Ave. Port Coquitlam, BC. Canada V3C 1T5. email mdangus@psychtest.com

Abstract

An appeal has been made by some educators over the past two decades to replace multiple-choice surveys of school skills achievement with extended-task performance-based (EP) items. Some jurisdictions have implemented EP-testing in their large-scale assessment. Issues are reviewed regarding validity and economics of large-scale provincial/state or national EP-assessment programs. EP-items tend to present difficulties in establishing accurate scoring criteria/rubrics, generalizability/accuracy problems, extended administration times, and onerous costs. To some extent, most of these difficulties have been overcome, except for costs. In general, it costs more to test one grade with valid EP-testing, than it does to test an entire school system at every grade with traditional multiple-choice tests.

The goals and costs of large-scale achievement testing.

The goal of survey testing of curriculum achievement is generally to establish how well students on an individual level are acquiring the curriculum AND in aggregate how well students within classes, schools, school districts or the larger school system are achieving curriculum objectives. For accountability purposes it is important that changes in performance of the school system over time be documented. This allows policy makers, teacher, parents and students to have a clear understanding of school and school system performance, and may also provide an objective source of useful information about individual learning. Several high-quality, commercially produced, multiple-choice-tests (MC-Tests), have been constructed for this purpose. In addition, MC-Tests have been developed by various ministries of education over the past several decades.

Recently a trend has emerged toward use of extended-task performance-based (EP) items, and also constructed response-items, in large-scale assessments. This has sometimes been justified using the theory that tests comprised of EP-items may be more valid than well-constructed MC-Tests. Yet an increment in validity for these types of tests has not been supported by research. Indeed, even with testing programs in place, good validity research in support of existing EP-testing programs is rarely conducted, perhaps because it adds to the costs of the programs.

Most arguments in support of EP-tests often seem to be an emotional appeal for authentic face-validity in the tasks students are required to undertake in an assessment. This is an appeal made by a great many educators and should not be dismissed out of hand. But no data has been forthcoming to support the hypothesis that task-authenticity automatically implies increased measurement validity. On the contrary, empirical data has frequently falsified this hypothesis (Shavelson, Baxter & Gao, 1993; also see the discussion of BC's *1998 Foundation Skills*

Assessment writing test, later in this paper). Instead research has shown that great care must be taken to avoid generalizability problems due to scorer error and other sources of error that are normally associated with “large person-task interactions *and* small numbers of tasks (Brennan, 2001)”.

Logically we must consider that EP-items tend to measure different things in addition to curriculum achievement. Frequently, in addition to curriculum knowledge these items require use of higher order thinking (Messick, 1996) in much the same way that is required on MC-tests of school ability (or IQ). Thus depending upon the item-design, use of EP-items may measure a construct that is more related to “G” than to curriculum-achievement. In addition, because scoring can be quite subjective, it is possible that EP- testing may forgo obtaining strictly objective information about individual student achievement, which is one of the inherent values of individual-level MC-test scores provided to the classroom teacher.

Economic considerations

Both validity *and* efficiency are necessary in large-scale surveys. By efficiency I refer to the concept of economic cost, which may be applied to all of the various cost-benefit tradeoffs which exist in large-scale educational testing.

MSPAP as an example of valid testing using EP-items

There is at least one high-stakes state-wide EP-only testing program in North America, for which validity studies have been performed. This is the Maryland School Performance Assessment Program (MSPAP, Rosenberger, 2000). This program focuses on providing information at the school and school district level in order to guide instruction, but does not collect scores which would be valid at the individual level.

In MSPAP, as elsewhere when EP-testing has been instituted, a political constraint was imposed that the testing program be an EP-design. It is well-known in economics that imposition

of design requirements for political reasons, which are unrelated to considerations of value, usually generates economic inefficiency. To show that this is the case, we may consider that if MSPAP is compared to other EP-only tests, it is actually a model of design-efficiency and great measurement validity. Yet when the constraint of using EP-items is removed, we see quite a different picture, at least in terms of economic cost.

MSPAP randomly assigns students in each school to receive different test forms which reflect different aspects of the State curricula. This is done in order to achieve comprehensive curriculum coverage at the school level, but not at the individual level. MSPAP takes 9 hours to administer, uses apparatus and non-print materials, and requires labor-intensive administration and scoring. Three *different* professionally designed test forms are used in every school. Three of these forms measure different aspects of the Maryland basic-skills curriculum. A fourth form

facilitates year to year comparisons. To obtain the same information with scores which would be valid at an individual level it would take *27 hours* of testing. Thus obtaining individual level scores using its EP-tests would cost Maryland approximately 3 times as much as obtaining the school-level scores it presently collects. In MSPAP, the cost-saving of not producing individual level scores is economically wise. The tests take so long to score, that by the time they are released during the following school year, the individual scores would only serve to audit teacher grades of the previous school year. Individual scores would have little or no educational relevance to students or teachers, and hence would have low economic value.

Opportunity costs of EP-testing may be calculated by comparing costs of a hypothetical expanded 27 hour version of MSPAP with an appropriate MC-Test battery which would provide valid individual, school, school district and state scores. The comparable MC-Test could certainly be administered in 5 hours of testing time and would be scorable using inexpensive computer scanning of student answer documents. Such MC-testing costs somewhere in the

neighborhood of \$5 to \$8 per student including scoring and reporting of results, and is currently available from several commercial test-publishers. Comparable EP-testing costs well-over \$100 per student.

The 27-hour hypothetical EP-testing would cost more than a traditional MC-Test by an administration-time factor of about 6 times as long, but by a dollar-cost factor of approximately 20 times as much. The higher total dollar-cost factor for EP-testing is due mostly to the costs associated with having human scoring personnel laboriously score each student's EP-task. To put this in economic perspective, consider that state-wide testing at a single grade-level each year with an EP-test, costs as much as testing every grade from K-12 every year, and sufficient resources would be left over to conduct appropriate validity study.

The technical quality of information obtained from EP-tests

This section discusses BC's Foundation Skills Assessment (FSA) Writing subtest as an example of a single-item EP test. This test was a 35 minute essay, based on a writing topic provided to students along with instructions. The task included 5 minutes planning prior to the essay writing and 5 minutes for proof-reading afterward.

The FSA testing being conducted in 2002 involves a separate focused writing passage of 15 minutes length, as well as the 35 minute extended essay. As well, longer planning activities, plus the opportunity to discuss with classmates before writing the essay, are provided. Reliability and validity data are not released by the Province any longer (J. Gaskill, February 2002, personal communication), which is why the present paper focuses on the results of 1998 testing.

Scoring reliability:

In the technical reporting of the 1998 FSA-writing test (Gaskill, 1999), it is stated that inter-scorer error was calculated by re-scoring nearly 4,500 student essays (about 1500 at each grade level in the testing). The percentage of agreements on the re-scoring study was reported along with correlations. Results seemed to indicate that scorers could not reliably assign the same score to a passage upon a second marking of the essay. Percentage agreement and

interscorer-correlation are

Table 1.

reproduced in table 1.

(source, Gaskill, 1999, pp.

38-43).

	Percent Agreement	Inter-marker Correlation
Grade 4	64.5	.56
Grade 7	53.5	.55
Grade 10	51.9	.57

The FSA technical report opined that the levels of scorer agreement reproduced here in table 1 were low, but still indicated adequate consistency in the scoring (p. 42-43).

Generalizability of the testing

We should not mistakenly confuse the interscorer correlation coefficient provided for the FSA-writing tests (Gaskill, 1999) with the overall reliability of the test-scores. Inter-scorer error on this testing compares with the error from scanning and scoring MC-tests which is near zero and should be compared to an interscorer correlation of 1.0. Overall generalizability/reliability of the EP-testing is the ratio of [(1- error variance) / Total variance]. This means that loss of generalizability due to all sources of error must be considered in the numerator of the reliability coefficient when arriving at an estimation of test reliability.

For EP-tasks reliability is often difficult to estimate, due to lack of statistically equivalent tasks, which would provide a method of estimating person-task variance, and due to the un-repeatable nature of some tasks. It has been shown empirically however, that if great care is not

taken, loss of generalizability may easily include 100% of test score variance (i.e. see Shavelson, Baxter and Gao, 2002). Once this happens, scores are meaningless.

The overall reliability of comparable MC-Tests tends to be around .8 for sub-scales and tends to be more than .9 for composite-scores (Nelson-Thomson, 1997; Canadian Test Centre, 2000). Overall reliability of FSA-writing cannot be calculated, but it is known to be substantially lower than .57 which is the reliability estimate that includes only error attributable to scoring.

Waiting time cost

In some senses test-score information is time-sensitive. FSA-writing test results are approximately three months old when they are delivered, which is a fairly normal delay that applies to nearly all large-scale EP-testing programs. In the FSA, teachers are provided with marking specifications and may collect unofficial scores for their students immediately. Official marking takes place over the summer, and results are returned to schools in the fall of the following school year. Three months is generally too long after testing for results to be of much educational use, however they are still of administrative use.

Conclusions

The discussion in this paper has shown that economic cost of conducting large-scale EP-testing on a province-wide, or national scale¹ is substantial. Such testing is being conducted in several jurisdictions. When testing programs are constrained by political considerations, the result should be expected to be either increased economic cost or lowered measurement-validity (value), or both. We see that this is the case. To identify true economic cost, the measurement enterprise should be analyzed by comparing it to the most cost-efficient valid alternatives –not merely by comparing to other similarly constrained programs.

The above discussion has implications for the calculation of indices of measurement value (validity). Measurement should be validated by comparing utility of reasonable methods of

assessing the same construct. Validity or reliability study should not merely be a same-method comparison between measures that were constructed in accordance with the same political, measurement-irrelevant assessment-method requirement. Such studies appear to affirm the validity of measures which would otherwise appear to be of questionable validity if they were compared fairly to the best alternative means of measurement (e.g. see the study by Wolfe, Wiley & Traub, 1999).

Reliable and valid testing using EP-items has been demonstrated in some jurisdictions such as the State of Maryland. The total cost including development, administration time and scoring, still are many times the cost of a valid commercially available MC-Test. Analysis here has also shown that there should be a further substantial discounting of economic return on EP-testing or testing which uses constructed response format items, due to costs associated with the lengthy time between assessment and delivery of test results.

Acknowledgements

I would like to thank Kadriye Ercikan and Monica Angus for comments and suggestions during the preparation of this paper.

References

- Brennan, R.L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*. 38 (4) 295-317.
- Canadian Psychological Association (1996). *Guidelines for Educational and Psychological Testing (First edition, 1987)*. Ottawa, ON, Canada: Author. Retrieved from the Internet, January 31, 2002. <http://www.cpa.ca/guide9.html>
- Canadian Test Centre (2000). *Canadian Achievement Tests and Tests of Cognitive Skills*, 3rd. eds. Scarborough, ON: Author.
- Cronbach, L.J., Gleser, G.C. Nanda, H. & Rajaratnam, N (1972). *The dependability of behavioral measurements*. New York: John Wiley & Sons.
- Gaskill, J (1998) *British Columbia 1998 Assessment of Reading Comprehension and First Draft Writing: Technical Report*. Victoria: British Columbia Ministry of Education.
- Messick, S. (1996) Validity of Performance Assessments. In *Technical Issues in Large-scale Performance Assessment*, Phillips, G. W. Ed. NCES 96-802.
- Nelson-Thomson Learning of Canada (1997). *Canadian Tests of Basic Skills, and Canadian Cognitive abilities Tests*. Toronto, ON: Author.
- Nelson-Thomson Learning of Canada (1981). *Canadian Tests of Basic Skills: Technical Manual*. Toronto, ON: Author.
- Otis, A. & Lennon, R. (1996). *Otis-Lennon School Ability Test*. San Antonio: The Psychological Corporation.
- Rosenberger, K. (2000). *The Maryland school performance assessment program (MSPAP)*. Maryland State board of Education. Retrieved from the internet March 16th, 2002. http://www.mdk12.org/mspp/mspap/how-scored/mspap_info/
- Shavelson, Baxter & Gao (1993). Sampling Variability of Performance Assessments. *Journal of*

Educational Measurement. 30 (3) 215-232.

Wolfe, R.; Wiley, D. & Traub, R. (1999). Psychometric Perspectives for EQAO: Generalizability Theory and Applications. *EQAO Research Series.* No. 3. Retrieved from the internet, March 22, 2002. http://www.eqao.com/eqao/home_page/pdf_e/99/99P033e.pdf

Footnotes

¹ There is at least one situation, in which testing based of EP items seems to have been economically valid. This is in multi-national assessment, which is not the main topic of the present paper. When surveys of curricula do not yield a broadly representative common curriculum – as in testing across many countries which perhaps have different cultures, languages, and different educational needs – development of valid testing may be extremely difficult and expensive. In this case it can be more economically efficient to use tests comprised of agreed-upon EP-items. This is a situation in which it is desired to take measurements and to compare relatively incomparable things. Agreed-upon EP-items can serve as proxies for actual curriculum achievement testing, with some anticipated loss in generalizability. That is, we must accept that such EP-tests probably do not measure curriculum-referenced achievement as we desire to do, but we will at least be able to make a comparison of sorts. EP-items generally require higher-order thinking, and so are more likely to be comparable to English-language MC-Tests of school ability (IQ), such as Test of Cognitive skills (Canadian Test Centre, 2000); Canadian Cognitive Abilities Tests, (Nelson-Thomson Learning, 1997); and the Otis-Lennon School Abilities Test, (Otis & Lennon, 1996). One could probably translate one of the available School-Ability MC-Tests or develop a similar MC-Test for the survey. However, differential item validity across language translations may be much easier and more cost-effective to manage using EP-format items in which much of the language used is in the item-prompt, or directions. In an EP science test for example, samples of dirt, beakers, graduates, testing solutions, and so forth, will tend to represent the same thing in all the languages of testing.