Robert J. Wilson
Queen's University

Aspects of Validity in Large-Scale Programs of Student Assessment

Large-scale programs of student assessment have increased in number and scope over the last two decades. Their approaches were largely derived from the technology developed for standardized achievement testing in the United States. Because large-scale assessment purposes are more extensive than testing purposes, the validity issues are more complex. This article explores what those differences are in such areas as item types, administration, interpretation, and standards. It concludes with some recommendations concerning the questions that need to be addressed if large-scale assessments are to accomplish their varied goals adequately.

Dans les vingt dernières années, les programmes à grande échelle visant le diagnostic des élèves ont connu une croissance, tant pour le nombre que l'envergure. Les approches sur lesquelles reposent ces programmes sont tirées, en grande partie, de la technologie développée pour les tests de rendement standardisés aux États-Unis. Puisque les buts des tests pour fins diagnostiques à grande échelle sont plus élaborés que ceux des tests pour fins d'évaluation, les questions de validité y sont plus complexes. Cet article repère ces différences dans des domaines tels le type d'items, l'administration, l'interprétation et les standards. La conclusion présente des recommandations quant aux questions qui doivent être abordées afin que les tests pour fins diagnostiques à grande échelle atteignent leurs buts variés de façon satisfaisante.

In Canada over the last two decades an increasing number of assessment programs have been dedicated to providing external accounts of student learning. Driven largely by political and economic forces, provincial, national, and some large local authorities have demanded and received reports of student achievement on key outcomes for their students. These key outcomes typically include the subjects of mathematics and the language arts and sometimes science and the social studies. Other subjects (or other ways of compartmentalizing outcomes) infrequently are the object of attention. It seems that a widely held belief among educators and the public is that numeracy and literacy are the foundation stones for building a learning edifice in students.

Centrality of Validity

The recipes for large-scale assessment were derived mainly from the technology surrounding standardized testing in the United States. The standardized testing technology of item selection, development, tryout, and analysis has been generalized to large-scale assessment and subsequently to classroom recommendations without much in the way of verification of the supposed connections to these different settings (Wilson, 1994).

Perhaps because so much of the technology of item analysis, item response theory, factor analysis, and standard setting is arcane to most educators,

Robert J Wilson is a professor of assessment and evaluation in the Faculty of Education.

politicians, and the public, it remains so unexamined. But the mechanisms exist in the assessment discipline itself for monitoring its effects, mechanisms that generally appear under the heading of validity. Validity is the *sine qua non* of assessment because it includes, and is affected by, the accuracy of the assessment as well as the purposes to which the assessments are put.

The issue of what constitutes valid measurement is the focus of much continuing discussion (Cronbach, 1989; Messick, 1989; Moss, 1992). One issue concerns the centrality of construct validation in the validity argument. A second issue concerns the degree to which the social consequences of a measure's use should be considered as part of the validity domain. Both of these issues are central to the following argument. Following authorities such as those listed above, I contend at least that content validity is an insufficient basis for assuming the internal validity of any achievement measure and that ignoring the uses to which the results of the assessment are put may also be shown to compromise the external validity of the results. Although these assertions are not as strong as these authorities and others might make, they are sufficient to my purpose in the following discussion.

Purpose

To organize what follows, I decided to highlight the processes of instrument development, administration, and interpretation used by specialists to accomplish their stated purposes. In describing what is typically done, I hope to show how validity is affected and therefore which claims are possible and which are not well supported by those processes. Finally, I provide some questions that seem central to valid assessments of the type described herein. Before looking at the technical aspects, however, it is important to review what we now know about the object of the assessments: learning.

Role of Learning Theory

Surprisingly, a theory of learning is not much discussed in the assessment literature. As others have pointed out, traditional assessment technologies grew up in a time when behaviorist theories of learning were dominant (Hager & Butler, 1996; Wilson & Kirby, 1994), and the procedures for much assessment practice still implicitly reflect that orientation.

Performance assessment, for example, provides a good illustration of this model of learning evident in assessment practice. In the Third International Mathematics and Science Study (TIMSS), for example, "applications to problems" were included as part of the general survey of learning under the heading of "performance assessment." Here is an example of one task from science and one from mathematics included in this assessment:

> Rubber Band: A rubber band with a hook on its lower end is fixed to hang vertically from a clip on a clipboard. Students measure the change in the length of the rubber band as they attach an increasing number of weights to the hook. Students record and tabulate their observations and then interpret them.
>
> Calculator: Students perform a set of multiplications with a calculator and observe and record patterns of results. These data allow students to predict the results of further multiplications beyond the scope of the calculator. (Martin & Kelly, 1998, p. 65)

As Roth and McGinn (1997) have shown, these types of school problems already have an answer implicated in the presentation, and students have the job of trying to determine what the problem statement hides, and then find what it was they were taught in school that might be brought to bear on the issue.

Other problems, of course, do not come with so much decided in advance. Yet to some degree this is not necessarily relevant to validity. In all sorts of situations there is no requirement that an instrument actually *appear* like the object of measurement. Lots of measures in life do not imitate the objects of measurement, but nevertheless give us good predictability about that object's characteristics. For example, most of us trust the red light appearing on the gas gauge to be a true indicator of a lack of fuel whether or not the gauge itself is pictured as a tank. But what this indirection in measurement does require is that the conclusions developed from the responses students make to these tasks actually reflect their learning, in this case their learning to solve mathematical and scientific problems. Or, to put it into a more likely technical scenario, that differences in students' success with these problems correlate with differences in their successful application of these skills to other types of problem-solving. This caveat is especially important, for example, if inferences about future success are to be made by decision-makers on the basis of these differences.

To follow the gas gauge scenario one final step, most of us also have faith that the red light coming on actually does indicate little fuel and that we had better take some action because of it. One of the legacies of the implicit behaviorist theory driving many assessment programs is that such a validity need seems irrelevant: the responses of students are considered to be direct evidence of the achievement of the construct itself. What we have learned in the last 30 years of cognitive and measurement research is that such an argument is unsustainable (Wilson & Kirby, 1994). At the front end of concerns about most large-scale assessment programs, then, is one that says that these programs may not be measuring what they say they are measuring in any modern sense.

Purposes of Large-scale Assessment

Whether measures should be as direct as possible depends to a large degree on the purposes to which the results are to be put. Although the dominant movement seems to be in the direction of more congruence between assessment and the classroom setting, some concerns accompany these trends.

Most large-scale assessment programs are aimed at more than curriculum implementation decisions. These additional purposes usually involve prediction: prediction of future success for students in higher education, prediction of employability, and prediction of economic well-being for society. Also explicit in most of the national and international studies is the desire to compare systems with one another. Although there does seem to be a need to highlight curriculum implementation, especially as this might lead to relationships among key variables affecting that implementation, these internal educational purposes do not seem to be the main interest, as indicated by the highlights of the reports (e.g., TIMSS, 1997).

One of the unstated and untested assumptions about large-scale assessment, by both supporters and funders, is that the knowledge, skills, and attitudes being measured—derived exclusively from set curricula—actually matter. Yet this is a critical assumption given the uses to which the results are frequently put. The fundamental validity issue is not so much to what degree students have mastered the set curriculum, but given the set curriculum and their understanding of it, what can students now be expected to do?

The validity questions are different in these two cases. In the decision-makers and funders' actions, it is clear that the issue is the degree to which the items on the student assessments ask students to display learning that matters in the world outside school; that helps students learn more complex material, maintain and grow in employment, and contribute to Canadian society. For educators, on the other hand, the issue is the degree to which the tasks match up with the intended curriculum set by the province or country. It is this goal that is most frequently attended to in large-scale assessments (probably because these programs are run by educators for educators no matter what the rhetoric) and by doing so the wider purposes lie, if not unattended, at least unexamined.

This narrower, educational view limits the scope of the claims that can be made. Even for those purposes of interest to educators such a limitation does not allow discussion of such things as the value of the curricula on which the tests are based. Nonetheless, the relevance and utility of the curriculum (untested) as well as its implementation (tested) are both key determinants of the success of school programs for students.

Item Types and Test Administration

Much of the design and administration of large-scale assessment are placed largely within the purview of assessment specialists, and first and foremost, they require reliable information. The cleanliness of the data-gathering allows for differences in scores to be attributed more often to the abilities measured by the test than to variations in test administration.

Several outcomes attend this goal of standardized administration. First of all, it introduces a systematic variation in the results that runs something like this: "This is how the children of School X in Province or Country Y performed on this assessment of Reading at the grade 6 level when children were regrouped according to age, required to sit in rows at their desks, worked silently and independently, and did all their work without their usual tools." Most often the conditional following "This is how the grade 6 children of School X in Province Y" remains understated in public reports. Several entities in North America (e.g., Maryland, Ontario, and Vermont) have attempted to change this by incorporating more ecologically valid approaches to assessing school outcomes. Some others (e.g., British Columbia and California) have retained or returned to traditional approaches. The costs involved in doing the former of course are, among other things, a decrement in the reliability of the scores. The costs of doing the latter are that more attention needs to be paid to the validity of the evidence.

A second feature of the concern with reliability requires student responses to be limited and conventional. Most of the testing has been done in the past with selection items, and even those items requiring more sustained responses have been scored using predetermined keys, another holdover from the behaviorist tradition. (A praiseworthy attempt by TIMSS to use trial data from students to create scoring rubrics for some performance items is an exception.) As a consequence, students may be unable to demonstrate their understanding of the issue when they have no opportunity to elaborate on their responses.

Interpretation

In the absence of any direct, empirically based correspondence between performance on the test and performance outside the test, results of these assessments must be interpreted in a purely judgmental way. Here, for example, is the modified Angoff approach chosen by the School Achievement Indicators Project. Panelists drawn from an array of educators and the interested public are provided with contextual information, sample responses, and demographics about the student population. They are then invited to set performance levels they would expect at various levels of the population. Instead of stopping at that point, and using a summary of the respondents' views on what would be acceptable attainment levels, what is called an "informed" review then occurs. There, with the guidance of educators (who frequently have a stake in the outcome and who have orchestrated the whole assessment exercise), the panelists are now given the actual results. What happens next is illuminating:

> Panelists then returned to small groups and were invited to provide commentary on the preliminary expectations. Once every panelist had spoken and group discussion was conducted, each panelist was given opportunity to privately revise his or her preliminary estimates in light of the actual results and the *insights* [italics added] generated in the small group discussion. (Council of Ministers, 1997, p. 65)

It is these consensus-seeking judgments that are then used as the reported standards. At the end of one of these exercises, panelists were surveyed about their reactions. Fewer than 3% rated the actual results as the main reason for the expectations they arrived at and "many panelists questioned the necessity of providing actual assessment results as evidence" (p. 68).

The panelists were probably reacting to the controlling environment in which they were asked to generate their standards. They were encouraged to set "realistic not idealistic" levels by looking at "actual samples of student work." When their individual judgments were made, they were then asked to engage in an "informed" review in order to "revise the preliminary estimates," helped along by facilitators who focused on the "mismatch between expectations and results."

How would this sort of activity relate to, let us say, a fire marshall's visit to my basement? Apparently, it would be all right for me to explain to the marshall that with preschoolers there's bound to be a mess near the furnace, that he ought to visit my neighbor's basement before condemning mine, and that we ought to have a discussion about "reality" before he gives an overall judgment about the degree of hazard involved in cardboard boxes piled beside the furnace!

Another critical difference between this type of judgment and that elicited through the Angoff method is that the fire marshall will require remedial action on those areas of weakness found. No such follow-up is required from the assessment program's endeavors. Any "improvements in curricula, instructional strategies, and public policy"—the goals of the overall program—are assumed to occur spontaneously from the exercise itself.

At play here is a critical difference between educational criteria and those available in other areas. There is no direct validity evidence for educational standards that links variations in present performance with variations in valued, external performance. The fire marshall knows that fires are likely where flammable materials are stored near furnaces. We are less sure about the relationships existing between variations in ability to "select the statement that gives the main idea of the story" and meaningful, extraschool activity.

Standards

The interpretation phase of the program ultimately depends for its success on working out what is meant by standards. No assessment program, whether teacher-developed or government-sponsored, can fully answer its mandate, or the demands of validity, without somewhere describing its standards. The Oxford Dictionary gives several definitions of the term, but the most general one says a standard is: "a definite level of excellence, attainment ... viewed as a prescribed object of endeavour." For teachers, the task of determining the "prescribed object of endeavour" is given in the curriculum they are required to teach. For assessment programs, however, the task of defining this aspect of standards is not so easily accomplished.

If they are not merely duplicating what teachers do in their classrooms every day, assessment programs have the further task of justifying the selection of what learning they are assessing on grounds other than "everyone else is doing something similar." Attempts by some interested groups in Canada such as the Corporate Council on Education (Conference Board of Canada, 1993) and in the US, the Department of Labor (Secretary's Commission on Achieving Necessary Skills, 1991) provide another, noncurriculum perspective on standards for public education. Even in the subject areas most frequently assessed by large-scale assessment bodies, there is discussion of what skills are most needed outside the classroom (National Council of Teachers of Mathematics, 1989). In all cases the argument is made that certain skills, attitudes, and values are critical to individual success and public well-being, and by inference that these ought to be the objects of educational effort.

The issue becomes more focused when another aspect of what is meant by standards is invoked. Standards, according to the Oxford dictionary again, also mean "the legal magnitude of a unit of measure or weight" and the "rule, principle, or means of judgment or estimation; a criterion measure." If assessment programs are to do something more than duplicate teacher judgment, then their criterion measures need to be of sufficient robustness that they can handle their application to such variables as curricular relevance, instructional efficacy, content coverage, and resource fit as well as the usual one of curriculum implementation.

There is some indication too that proponents of large-scale assessments actually believe their tools would help in this regard. Here, for example, is the introductory statement to the TIMSS Technical Report:

> The ultimate goal [of TIMSS] is to isolate the factors directly relating to student learning that can be manipulated through policy changes in, for example, curricular emphasis, allocation of resources, or instructional practices. (Martin & Kelly, 1998, pp. 2-3)

For the School Achievement Indicators Program (SAIP), the program is supposed to "suggest improvements in curricula, instructional strategies, and public policy" (Council of Ministers of Education, Canada, 1997, p. 63). As soon as one asks questions beyond implementation—and both these programs as well as most others claim to—the discussion over standards immediately begins to transcend narrow instructional goals and moves into broader societal expectations.

The SAIP provides a good illustration of the issue even if it is not alone in that. SAIP assumes that the written documents, called the "intended curriculum," actually reflect what society wants its students to learn. How society would determine this without evidence of what students presently can do, cannot do, and should be able to do—in other words, the kind of data assessment programs are capable of providing but do not—is left unexamined. I am not suggesting that SAIP define those valuable goals, only that they contribute information that would inform the discussion.

Here is where the implicit behaviorism of most assessment models comes back to haunt their proponents. They believe that the responses students give to the selection items and the few supply items are inherently valid if they look like questions the students have already practiced in school or are easily extendable to such questions. The issues of whether it *matters* that students can or cannot do those questions, whether another set of questions entirely would be much more critical to success for these students and for society, does not enter the equation.[1]

How would one go about finding out whether another set of questions would be more powerful? It requires that a formal analysis be completed of what types of mathematics (or science, or reading, or writing) are actually in use throughout society at all levels, of what trends are evident in the field and in practice, and what values might be enlisted to prepare students for the future they will inhabit. One way to help with all of these questions would be to provide relevant data about what students can presently do on questions that predict such aims.

At present, the decision-makers who finance the study, and who wish to use its results for public policy, assume that the outcomes tested were worth achieving, that they should matter to students to have achieved them, that society benefits from students having achieved them, and that those students who score well on these tests will be more likely to succeed as adults, particularly in the economic race. One deleterious effect of such reasoning is that it is usually teachers and principals who are admonished to work harder to improve the results, whereas those responsible for supplying them with relevant curriculum and current resources for implementation escape accountability because their contribution is not scrutinized by the test.

We have recently witnessed the potential fallout of this approach in Ontario. Here a government committed to improving the results of Ontario's students on large-scale assessments like TIMSS and SAIP mandated a province-wide testing system geared to specific learning outcomes that themselves had to be manufactured quickly by the Ministry of Education with the curriculum documents to go along with them. Ironically, the government will soon be able to announce higher achievement for students in Ontario. What this means, in fact, is that Ontario teachers now will teach more like each other, test like each other, have their students learn like each other, and develop in them more of the skills tested by the test than they did before. Homogeneity in the school system will be equated with higher achievement. The fundamental validity question—Does it matter that they do better on these tests?—remains unanswered.

Handling the Validity Issue

A starting point for resolving the differences in achievement and validity evidence different clients need to make their various claims would be to focus not on the instruments, but on the descriptions of competence used to assess achievement. Thus far educational standards involve descriptions of different levels of curriculum implementation evidenced in the students' responses. (To see further evidence of the weakness of even that claim, see Nuthall & Alton-Lee, 1995, and their empirical investigation of the sources of student responses to assessments of curriculum-embedded learning.) It may be possible, however, to describe levels of attainment that honor both the curriculum that was implemented and the skills needed to participate effectively in extraschool behavior. This will involve a careful attention to magnitude descriptions.

Magnitude issues involve specifying different levels of attainment possible along a standard dimension. For problem-solving, for example, it is important to state not only what the standard is, but also what represents movement toward the ideal. For many of those involved in mathematics, for example, this dimension involves an increasing ability to solve text-based, one-step, two-step, and multi-step problems. For teachers this may be the best approach to aiding both them and their learners in the problems of instruction (see the teachers' reports in Ross, McKeiver, & Hogaboam-Gray, 1997 for an example of this application). For large-scale assessment designers, however, this tack is clearly wanting. What some of their clients wish to know are answers to questions like the following.

1. How able are our students in handling real-life problems?
2. Are students able to recognize situations in which mathematical solutions are possible and then apply mathematical principles to them?
3. Do these levels of achievement describe levels of success in nonschool activities involving mathematics?
4. Are any skills known that would be useful but are not present in the current curriculum?

The rubrics needed by those involved with this level of assessment are those that clearly apply not only to the items in question, but also to the prediction of levels on attributes that were not assessed directly. Notice that one could replace the term *mathematics* in these statements with *language skills* and the same principles would apply. Once again, the developments in cognitive theory might be good places to begin the search for more generalizable answers. Biggs and Collis (1982) with their SOLO taxonomy and Wilson (1996) with his

ICE model are examples of this less restrictive approach. Using nonspecific rubrics would free the developers from a dependence on specific items and encourage the specification of more generalizable data.

Standard setting activity could then be used to describe the acceptability of the measured levels of attainment independent of the particular items that were used to measure them. This would encourage the minimization of the role of educators as well, those who have a vested stake in the outcomes. After all, the first principle of audits is that the subjects of the audit do not conduct it. To inform this kind of decision, empirical data linking performance and future practice would be indispensable.

The principle behind the establishment of standards is that they represent the views of the community (however narrowly or widely defined) both on what is important to know and what represents the expected degrees of attainment. All this must be done before data are presented on what the current cohort is able to do. Otherwise the standards themselves become the focus of controversy and debate rather than the achievement levels of the students on them. Kane (1994), in his review of standard-setting for performance measures, put it this way:

> The collection of input from stakeholder groups is a time-honored part of democratic processes for establishing public policy and can therefore be considered a reasonable way to support the appropriateness of the policy decision to set the standard at a particular level. (p. 454)

Only when the dual process of standard development followed by matching to attainments is completed can attention be paid to what to do about instruction, the curriculum, resources, and other key educational variables. A valuable role for large-scale assessments, then, is to provide data to decision-makers that would help them with these questions. As presently constructed, following a minimalist content validity paradigm, they are unable to do so.

Conclusions

Both large-scale assessment practice and school practices have sprung from a behaviorist tradition represented most forcefully by standardized testing practice. This tradition is under attack from many sides. Many teachers, especially at the elementary level, have abandoned the static cognitive models of the past in favor of more dynamic ones. Large-scale assessment programs will have to adapt too if they are to remain viable. If the analysis above is accurate, then the following represent some of the key questions that need to be addressed by advisors to such programs.

1. Are the outcomes being assessed worthwhile and justified as such?
2. Are the standards (including overall criteria as well as their levels) well accepted by more than educators?
3. Do the tasks selected fit with these outcomes and standards?
4. Are students able to demonstrate their knowledge under representative conditions?
5. Is the scoring done reliably enough for group comparisons?
6. Does the magnitude of the differences reflect important dimensions related to standards?
7. Is the interpretation of the results trustworthy and defensible?

8. Do the recommendations flow logically from the results?
9. Do these recommendations apply to more than curriculum implementation?

Whether or not the interest in school achievement continues to grow exponentially as its relationships to economic well-being become more widely accepted, there is a need for assessment specialists to reexamine their own assumptions. In the early days of such assessments, the model used was the one that came readily to hand. In these latter days, this model is dated and inadequate. Classroom teachers are themselves beginning to alter their models to fit the more dynamic view of learning we now know is more valid. This trend needs to be encouraged by measurement specialists. Perhaps it is time those of us with understanding of the conceptual and technical issues involved in large-scale assessment also began to tend our own undernourished garden.

Note

[1] Phil Nagy (personal correspondence) has pointed out the salient fact that many of the panelists, successful adults all, who are called on to judge the success of mathematics achievement of 12- and 16-year-olds are frequently unable to complete the items successfully themselves.

Acknowledgments

I am indebted to John Anderson, Phil Nagy, David Robitaille, Todd Rogers, and Alan Taylor for their comments on an earlier draft of this article. They are, nevertheless, not responsible for the ideas expressed herein.

References

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning.* New York: Academic Press.

Conference Board of Canada. (1993). *Employability skills profile: The critical skills required for the Canadian workforce.* Ottawa, ON: Corporate Council on Education.

Council of Ministers of Education, Canada. (1997). *Technical report: Science assessment: School achievement indicators program.* Toronto, ON: Council of Ministers of Education.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.E. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

Hager, P., & Butler, J. (1996). Two models of educational assessment. *Assessment and Evaluation in Higher Education, 21,* 367-378.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425-461.

Martin, M.O., & Kelly, D.L. (1998). *TIMSS technical report, Volume 1: Design and development.* Chestnut Hill, MA: TIMSS International Study Center.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229-258.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal, 32,* 185-223.

Ross, J.A., McKeiver, S., & Hogaboam-Gray, A. (1997). Fluctuations in teacher efficacy during implementation of destreaming. *Canadian Journal of Education, 22,* 283-296.

Roth, W-M., & McGinn, M.K. (1997). Toward a new perspective on problem solving. *Canadian Journal of Education, 22*(1), 18-32.

Secretary's Commission on Achieving Necessary Skills. (1991). *What work requires of schools: A SCANS report for America 2000.* Washington, DC: US Department of Labor.

TIMSS. (1997). *Mathematics achievement in the primary school years.* Chestnut Hill, MA: TIMSS International Study Center.

Wilson, R.J. (1994). *Back to basics: A revisionist approach to classroom-based assessment.* Invited presidential address to the Canadian Educational Researchers' Association Annual Meeting, Calgary.

Wilson, R.J. (1996). *Assessing students in classrooms and schools.* Scarborough, ON: Allyn and Bacon.

Wilson, R.J., & Kirby, J.R. (1994). Introduction: Special issue on cognition and assessment. *Alberta Journal of Educational Research, 40,* 105-109.