REFLECTIONS ON THE USE OF LARGE-SCALE STUDENT ASSESSMENT FOR IMPROVING STUDENT SUCCESS

Charles Ungerleider

In 2003, I published an article setting out conditions that I believed should be met if large-scale student assessments would be useful for improving student achievement (Ungerleider, 2003). My purpose in writing the article was to provide policymakers with suggestions for maximizing the benefits of using such assessments. I argued that, at a minimum, the following conditions (after Willms, 1998, 2000; cf., Woessmann, 2001) must be fulfilled by any jurisdiction contemplating the use of large-scale student assessment to improve student achievement:

- establish broad agreement about what school outcomes are essential for all students;
- ensure that these areas are clearly articulated in the curriculum and are supported with appropriate instructional material;
- hold students, parents, and teachers accountable for those outcomes;
- assess student progress in the areas of importance at different times over their school careers;
- prepare teachers and encourage them to use teaching strategies that increase learning outcomes for all students;
- encourage mixed-ability grouping and discourage grouping, tracking, or streaming students by socio-economic background or in ways that increase differentiation among students of different ethno-cultural backgrounds;

CANADIAN JOURNAL OF EDUCATION 29, 3 (2006): 873-883

- assess schools on the basis of student growth in learning outcomes, taking into account their individual socio-economic backgrounds, the socio-economic context of the school community, as well as school policies and practices known to influence the achievement of the valued outcomes;
- examine rates of student progress as well as gradients in student progress associated with such background factors as socio-economic standing, gender, and ethnicity;
- ensure that teachers and administrators are well prepared for their responsibilities;
- counter misuse of the results of large-scale assessments in the media and elsewhere; and
- provide teachers with adequate time to individually and collectively interpret data for the purpose of improving instruction.

The suggestions I made were intended as a complement to the useful Principles for Fair Student Assessment Practices for Education in Canada (http://www.bctf.bc.ca/education/assessment/FairStudentAssessment.pdf#search=%22principles%20of%20fair%20student%20assessment%20in%20Canada%22), which also addresses the uses and limits of large-scale assessments and the conditions that must be in place for their results to be used in alternate contexts.

Between the time that I wrote the original article and the present, I enthusiastically took part in a research program as a member of the group of researchers whose work appears in this issue, participated as a member of a team that undertook analyses of the data from the 2003 mathematics assessment conducted under the banner of the Programme for International Student Assessment by the Organization for Economic Cooperation and Development (http://www.pisa.oecd.org/pages/0,2987,en_32252351_32235731_1_1_1_1_1,00.html), and remained an avid observer of the laudable efforts of provincial ministries of education to improve student achievement with the support of information provided by large-scale student assessments.

These experiences have prompted me to reconsider the use of large-scale student assessment to improve student academic achievement. The

more I am involved with such assessments, the more concerned I become.

The validity of any assessment result is conditional on the fit between the purpose for which the assessment was designed and the use of the results. At present there is no established procedure evaluating the validity of the alternative uses to which large-scale assessments might be put. It is not unusual that an assessment developed for one purpose to be used for another. At present, the developers of the original test leave it to subsequent users to decide if its use is appropriate.

I am concerned, too, with transforming large-scale assessments designed for certification or inter-jurisdictional comparison to improve student achievement. Most system-wide assessments that provincial governments employ were developed to certify the achievement of students.

I am increasingly of the view that such assessments are mismatched with the objective of improving student outcomes. System-wide assessments are too narrowly focussed in at least two significant ways. First, they do not adequately assess the broad range of knowledge and competencies that schools are intended to develop. For example, art, music, citizenship, or social responsibility outcomes are not assessed. Second, system-wide assessments are insufficiently attentive to the breadth of knowledge in the domains they do address. For example, in language arts, the appreciation of literature, oral fluency, and listening are often missing from such assessments, but are nonetheless important goals of the discipline.

The relationship between curriculum and assessment should require that all domains that are in the curriculum should be assessed, and that all things assessed should be in the curriculum. Where teachers are responsible for the whole curriculum but are held accountable only for selective portions of it inevitably encourages resentment in teachers and discourages instruction in those subjects and domains not represented on the assessments.

Part of the problem is that – as in almost all areas – financial considerations drive provincial, national, and international assessment programs, limiting educational and professional considerations. It would

be desirable, but costly, to increase the domains addressed by large-scale assessments to include those not presently represented. But, the inclusion of these domains would counter the criticism so frequently levelled that schools (and society) do not value what they do not formally assess.

It is equally desirable to broaden existing assessments of reading, writing, mathematics, and science to appraise dimensions of student performance that are not presently included. Financial considerations limit the opportunities that students have for representing their knowledge (in oral presentations or using graphics, for example).

It would be helpful to engage teachers working on the same staff or in the same subject areas in the discussion of questions used on the assessments. However, some jurisdictions are unwilling to provide the questions for this purpose, wishing to maintain secrecy. One of the main reasons is that it is costly to create good questions, and making them available would necessitate the additional expense of devising multiple versions of the assessment tool.

Another set of conditions limiting the utility of large-scale student assessment for the purposes of improving student performance pertains to system capacity. Under this rubric, I include a number of issues and problems. For example, I am concerned about the fairness of the assessments used. I wonder whether students in rural areas interpret the items on the assessments in the same way as urban students. The same concern applies to the interpretive frameworks of Aboriginal and non-Aboriginal students; boys and girls; immigrant and native-born students; and minority and majority language groups. I believe that jurisdictions have an ethical and professional obligation to determine whether the different populations to which system-wide assessments are administered perceive and respond to the tests in the same way. The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question.

Although it is possible to use a variety of analytical techniques to determine whether different categories of youngsters interpret and respond similarly to the same items, it takes time and financial resources

to carry out such analyses. Few jurisdictions systematically analyse the fairness of their assessments in this way.

I am also concerned about the misuse of the data from large-scale student assessments, and especially about comparing schools with one another. To be effective, assessment programs must link assessment data to information about policies and practices that are amenable to influence through the decisions that are made. The practical limits to improving student learning outcomes includes recognition that in Canada approximately 70 per cent of the variation in student learning is not attributable to school factors but to student, family, and community characteristics. To put it another way, the cumulative impact of school factors in student achievement is typically less than 30 per cent (see, for example, Anderson, et. al, 2006). Most of the variation attributable to these factors occurs *within* rather than *between* schools. This fact has escaped the attention of the public, politicians, and even education practitioners who invest importance in the crude and misleading comparisons of schools undertaken by organizations such as the Fraser Institute and the Atlantic Institute for Market Studies.

Nonetheless, 30 per cent is a significant amount of variation which, if understood and amenable to influence, has the potential to substantially improve educational outcomes. There are a variety of things one must have or do to maximize the opportunity to affect change. First, and most important, one must possess relevant, timely, and systematically gathered quality information about schools and students' prior instruction. The difficulty lies in knowing what information is likely to be important. Although the cumulative body of research literature has proved useful, it does not provide sufficient guidance. When variables identified in the literature are used, there remain large amounts of unexplained variation in student achievement.

Supplementary data collection is another multidimensional area of concern. Most large-scale student assessment programs do not systematically collect and analyze information about student, school, or home variables that could have substantial relationships to student achievement. Nor is adequate provision made for integrating available administrative information with the assessment data. Collecting such data routinely and ensuring its integration with the results of

assessments is costly, and few jurisdictions have willingly invested in the necessary infrastructure.

Few jurisdictions have made sufficient provision for the kind of analysis of the results of system-wide assessments that is needed to ensure the availability of information to support decision making, including linking the results with the aforementioned supplementary data. Such work is expensive, requiring expertise that is not in great supply.

Unless there is an active research agenda meaningfully attached to every assessment program, the research generated will be sparse and very slow in becoming available. Reliable data collection is a significant impediment to the development of such research. To the extent that students are the main source of information about instructional practices, the data are likely to be of poor quality. The same is true of reports about instructional practices made by administrators.

Careful observation of teachers at work may provide useful clues to the sorts of instructional practices worth investigating. But, before codifying them in teacher questionnaires, they should be thoroughly established as key correlates of learning outcomes and comprehensively discussed with experienced practitioners.

Even if the aforementioned factors were adequately addressed, I would continue to be concerned for several additional reasons. Many schools have significant numbers of transient students. Annual snapshots of student performance will be inversely accurate to the extent of student migration. A better approach is to monitor student trajectories over time, a time consuming and expensive undertaking.

To be effective in improving student success in school, the results of assessments must be communicated to relevant audiences in a timely and thoughtful manner. Newfoundland makes provision for ministry personnel to visit schools and help the staff to interpret results; but most provinces do not (G. Galway, personal communication, November 9, 2005).

Two obstacles stand in the way of the effective use of the results of system-wide assessments. First, those who might use the results to inform their decisions often do not know how to do so and, even in those instances where such capacity exists, the opportunities for making use of

the information are too limited. For example, there is too little time for school staffs to interpret the information pertinent to their setting and to consider how the information might be used to inform their decisions about policy and practice.

The current atmosphere surrounding the use of system-wide testing for student improvement is a second obstacle. Many teachers are not receptive to system-wide testing regimes. Teachers have a variety of legitimate concerns about such regimes that in most jurisdictions have not been adequately addressed. These include concerns about the focus of the tests, their validity and reliability, the interpretation given to the results, and the implication that the teacher in whose class the test was administered is primarily responsible for the student outcomes. Jurisdictions must recognize the centrality of teachers to the process and the importance of winning teacher acceptance of large-scale assessment programs.

Teacher cooperation is integral to acquiring useful information about instructional and school practices, to the interpretation of results, and to the planning and implementation of policies and practices that may improve instruction. Unless there is substantial improvement in the atmosphere surrounding system-wide testing regimes, teachers are likely to remain suspicious about them. They will be unwilling to lend their considerable knowledge and expertise to the improvement of the testing regimes. Accountability regimes in jurisdictions with well educated teachers such as those found in Canada must be predicated on enabling teachers rather than controlling or "fixing" them.

Despite my pessimism about the use of system-wide assessments for improving student achievement, there are some hopeful signs. Ontario has recently made several efforts to address this challenge. The Ontario Ministry of Education has undertaken an initiative to provide support for building capacity in local school boards for "Managing Information for Student Achievement" (MISA) (https://imgdata.edu.gov.on.ca/MISADOCS/Misa.html). Recognizing that individual schools might not have the capacity to address needed improvements, the Ministry also provides "Turnaround Teams" that include "experienced principals, expert teachers and external literacy experts who provide support to improve the acquisition of literacy by students in Junior Kindergarten

through Grade 3" (http://www.edu.gov.on.ca/eng/teacher/help.html). The teams are sent to schools to "identify strengths and weaknesses in the school's instructional practices; share successful instructional approaches, assessment and leadership strategies; and provide ongoing mentoring and support that addresses the unique challenges of each school." Ontario's approach is not based exclusively on the assumption that, by focusing on instruction, the schools will improve literacy levels. Other school-related interventions (for example, altering class sizes, changing school discipline policies, improving attendance, creating a supportive atmosphere) are also recognized as being needed to improve performance and sustain it over time.

Ontario is not alone in contemplating how assessments can contribute to school improvement. British Columbia has been generous in making its assessment data, along with other school and student related information, available through Edudata Canada (http://edudata.educ.ubc.ca/), an independent organization that facilitates access to such information for qualified researchers. British Columbia is introducing a new information management system, British Columbia Enterprise Student Information System (http://www.bcbudget.gov.bc.ca/Annual_Reports/2005_2006/educ/educ.pdf), that holds promise to provide the kind of supplementary information I have indicated might prove helpful in identifying factors amenable to influence. However, teachers are likely to resist its implementation because they feel it is being undertaken without their involvement.

Since 1999, Alberta has pursued school improvement under the auspices of the Alberta Initiative for School Improvement (AISI) (http://www.education.gov.ab.ca/k%5F12/special/aisi/), providing resources for school communities to engage in a wide variety of disparate initiatives to improve schooling. Manitoba has been engaged in similar efforts for some time through its Manitoba School Improvement Program Inc. (MSIP) (http://www.msip.ca/), "an independent, non-profit, non-governmental intermediary school improvement organization dedicated to supporting youth through the improvement of public secondary schools in Manitoba."

As mentioned earlier, the expertise required for analyzing the results of system-wide assessments is in short supply and very costly. As a consequence, the Canadian Council on Learning (http://www.ccl-cca.ca/ccl) is developing an Internet-based tool to enable those responsible for assessment at the classroom, school, school board, or provincial to carry out such analyses.

With universal accessibility of computers, there is no longer a need to use standardized tests to measure performance on a standardized scale. Given a common bank of test items, tools such as the one being developed by the Canadian Council on Learning would allow teachers to create classroom assessments that suit the needs of their students while still benchmarking their students' performance against provincially established standards. I believe that this approach would encourage a better fit between assessment practices and the curriculum.

There can be little doubt that building the capacity for using assessment to improve student success is a long-term project. It will take between 10 and 20 years to fulfill the conditions and overcome the obstacles that I have enumerated here. Moreover, there are additional impediments to overcome. Chief among them is that education does not have a record of sustaining effort toward long-term goals. The temporal horizon of most politicians and political parties is considerably more limited than the time required. Regime changes typically signal changes in direction or emphasis. Nonetheless, the use of large-scale student assessment as a tool for improving student achievement is one that parties across the political spectrum have embraced and sustained even in the face of significant political changes such as have occurred in Ontario and British Columbia.

Another impediment is the significant investment that will be required to develop the necessary infrastructure for collecting and analyzing information about students, school policies, and instructional programs. But that infrastructure will be needed for identifying factors amenable to influence that are related to student achievement. Furthermore, additional resources will be required to free teachers from some of their responsibilities so that they may work together with administrators at the school level to connect their understanding of the data to the decisions made about policy and practice.

Faculties of education must recognize their obligation to prepare teachers with the dispositions and knowledge they will need to operate in a changed milieu. This will require collaborative effort among professors, who typically do not communicate with one another. A useful starting point for such a discussion might be: What do beginning teachers need to know about valid assessment practices and evidence-informed decisionmaking?

Ultimately, the probability of success will be increased if the project is understood as being long term, requiring continuing political commitment and significant resources. One additional change that is required, but does not depend on significant expenditure of financial resources, is the creation of a sustained atmosphere that is respectful of teachers and their professional judgment. This above all will be necessary to make large-scale student assessment a useful tool to improve student achievement.

ACKNOWLDGEMENTS

I am grateful to the helpful suggestions of a number of people, including John Anderson, Fernando Cartwright, Kit Krieger, Sam Robinson, and Mary Ungerleider—none of whom is responsible for my comments.

NOTES

[1] I might feel better about such transformative uses if the original developers assumed responsibility for providing an evaluation of the appropriateness of the use. They might, for example, assign a value between 0 (inappropriate) and 10 (appropriate) for appropriateness of use. In the case of using large-scale provincial assessments for ranking schools without adequate statistical control or theoretical justification, the evaluation would probably be around 1.

[2] The Internet-based tool automates test construction from an item bank; scores test data using user-supplied answer keys; uses plain language to explain how items need to be improved to make the final test more efficient and effective; illustrates statistical results graphically; provides easy-to-follow instructions and comprehensive error handling; uses standard naming conventions and file formats to maintain data; calculates classical item and test statistics for use with pilot test data; gauges the extent to which items are measuring what they ought to be (factor analysis, item dependency analysis, and item bias analysis); uses Item Response Theory (IRT), item calibration, and score

estimation; estimates optimal cut-points for reporting of results by proficiency levels (for norm or criterion-referenced assessments); employs IRT test equating for assessment programs that use common items across exams; generates efficient tests from pre-calibrated item banks; reads EXCEL, delimited, or SPSS formatted files; and converts results to EXCEL, *.csv, text, or SPSS formats. The tool is currently being subjected to a variety of tests and will be field tested during 2006-2007.

REFERENCES

Anderson, J. O., Rogers, W. T., Klinger, D.A., Ungerleider, C., Glickman, V., & Anderson, B. (2006.) Student and School Correlates of mathematics achievement: Models of school performance based on Pan-Canadian student assessment. *Canadian Journal of Education*, 29(3), 00000

Ungerleider, C., (2003) Large-scale student assessment: Guidelines for policy-makers. *International Journal of Testing*, 3(2), 119-128

Willms, J. D. (1998). Assessment strategies for Title I of the improving America's Schools Act. Report prepared for the Committee on Title I Testing and Assessment of the National Academy of Sciences.

Willms, J. D. (2000). Monitoring school performance for "standards-based reform." *Evaluation and Research in Education*, 14(3&4), 237-253.

Woessmann, L. (January, 2001). *Schooling, resources, educational institutions, and student performance: The international evidence*. Kiel, Germany: Kiel Institute of World Economics.