The Canadian Journal of Program Evaluation Vol. 22 No. 3 Pages 53–71
ISSN 0834-1516

DESIGN AND DEVELOPMENT ISSUES IN PROVINCIAL LARGE-SCALE ASSESSMENTS: DESIGNING ASSESSMENTS TO INFORM POLICY AND PRACTICE

Kadriye Ercikan
Stephanie Barclay-McKeown
University of British Columbia
Vancouver, British Columbia

Abstract: Over the past four decades, there has been much debate on key sources of data in evaluating education, determining school effectiveness, and providing evidence to inform accountability and education planning. Entangled in this debate has been the extent to which large-scale assessments of learning provide valid evidence about the quality of schooling and education in Canada and how they can be used to inform education practice and policy. This article discusses five issues in large-scale assessments that are key for their usefulness and for making valid inferences. Based on recent research on assessment design and validity, the authors offer recommendations for large-scale assessments to better serve the multiple purposes they are intended to serve.

Résumé : Pendant les quatre dernières décennies, il y a eu bon nombre de discussions autour des bases de données créées pour l'évaluation de l'éducation et de l'efficacité des écoles ainsi que de leur capacité à fournir des données probantes pour la reddition de comptes et la planification de l'éducation. Le débat incite à se poser plusieurs questions dont : « Dans quelle mesure les évaluations à grande échelle sur l'apprentissage offrent-elles des données concrètes et valides sur la qualité de l'éducation au Canada? » et « Comment ces données peuvent-elles être utilisées en vue d'éclairer les décisions pratiques et politiques? ». Cet article porte sur certains aspects à prendre en considération dans les évaluations à grande échelle afin de tirer des conclusions valides. En s'inspirant de recherches récentes sur le développement et la validité des évaluations à grande échelle du rendement des élèves, les auteures offrent des recommandations par rapport à l'élaboration de telles évaluations qui permettront d'accomplir plus efficacement les divers usages pour lesquels elles ont été développées.

Corresponding author: Dr. Kadriye Ercikan, Faculty of Education, Educational and Counseling Psychology and Special Education, University of British Columbia, 2125 Main Mall, Vancouver, BC Canada V6T 1Z4; <kadriye.ercikan@ubc.ca>

In reaction to multiple demands on existing resources, evidence-based decision making for evaluation and planning has become the norm for public education in many countries, including Canada (Crundwell, 2005; Earl, 1999; Stone, 1999). Over the past four decades, there has been much debate on key sources of data in evaluating education, in determining school effectiveness, and for providing evidence to inform accountability and education planning (Crundwell, 2005). Entangled in this debate has been the extent to which large-scale assessments of learning provide valid evidence about the quality of schooling and education in Canada and how they can be used to inform education practice and policy (Chudowsky & Pellegrino, 2003; Ercikan, 2004, 2006a; Gao, Shavelson, & Baxter, 1994; Goldstein, 1997; Leighton, 2004; Linn, 1993; Linn, Baker, & Dunbar, 1991; Yang & Goldstein, 1999). Provinces have been responding to this climate of accountability, and all provinces in Canada currently conduct large-scale assessments of learning at elementary and secondary levels of education. In Canada, similar to many other countries, multiple stakeholders are involved in public education, and consequently large-scale assessments have been intended to address a variety of needs of the stakeholders. Practitioners would like to use test results to inform classroom instruction and school policy, while the public wants to know that their tax dollars paid to education translate into student learning, and ministries want assurances that students are achieving a certain standard of learning (Crundwell, 2005; Earl, 1999; Taylor & Tubianosa, 2001).

Most Canadian provinces have been relying on the results of these assessments to determine a wide range of critical decisions including student access to educational scholarships, graduation from high school, and entrance into post-secondary educational institutions (Crundwell, 2005; Earl, 1999; Taylor & Tubianosa, 2001). For example, in British Columbia (B.C.), students are required to write a minimum of five provincial examinations to graduate with their results counted toward their final grades, which is taken into consideration when awarded the Dogwood diploma (e.g., graduation), scholarship awards, and for application to B.C. universities (B.C. Ministry of Education, 2007).

Even though large-scale assessment results have been used for a variety of decisions in most provinces in Canada, evidence of reliability and validity of these assessments is rarely reported and the assessments themselves often do not fulfill the accepted test development standards (Crundwell, 2005). The purpose of this article is to contribute to the ongoing discussion on designing and developing large-scale

assessments in order for them to meet some of the key purposes they are intended to serve. The article intends to draw attention to key large-scale assessment issues in the Canadian context and provide guidance to assessment design and development issues.

First, we summarize and discuss two aspects of large-scale assessments that have been the focus of major criticism and discussion. One is the extent to which the assessments are connected to school learning and the extent to which they inform school learning. The second set of criticisms and discussions has been related to the use of assessment results in accountability models. Following discussion of these two issues, we review the goals and purposes of provincial large-scale assessments. The next section focuses on discussing five key design requirements pertinent to all large-scale assessments intended to assess learning. These requirements are related to (a) importance of the types of learning outcomes included in assessments; (b) alignment with learning, instructional, and curricular goals; (c) measurement accuracy; (d) score reports; and (e) consequential aspects of large-scale assessments. Based on recent research on assessment design and validity, we offer recommendations for the implementation of these design requirements into current large-scale assessment programs (Black & Wiliam, 1998; Chudowsky & Pellegrino, 2003; Crundwell, 2005; Ercikan 2006b; Leighton, 2004; Linn, 2000; Nichols, 1994; Schafer & Moody, 2004; Shepard, 2000; Taylor & Tubianosa, 2001).

CONNECTIONS WITH LEARNING

The increased demand for assessment information and dedication of resources to assessing students for accountability at the system level have been accompanied by a growing awareness of the need for assessments to have a stronger connection to learning so that they can be used appropriately to inform classroom practice and school and district policies related to learning (Chudowsky & Pellegrino, 2003; Ercikan, 2006b; National Research Council, 2001; Sicoly, 2002). The reactions to the increased levels of assessment for accountability purposes as well as progress in our understanding of how students learn have contributed to this trend (Ercikan, 2006b; Leighton, 2004; Mislevy, Wilson, Ercikan, & Chudowsky, 2002; National Research Council, 2001; Nichols, 1994). Educators and educational assessment experts have provided evidence and convincing arguments that better connections between assessment and instruction can have a positive impact on learning (Black & Wiliam, 1998; Chudowsky & Pellegrino, 2003; Ercikan, 2006b; Gipps, 1999; National Research Council, 2001;

Shepard, 2000). Currently, there is little evidence of these connections in large-scale assessments in Canada or elsewhere (Carnoy, 2005; Chudowsky & Pellegrino, 2003; Ercikan 2006b; National Research Council, 2001).

USE OF ASSESSMENT RESULTS IN ACCOUNTABILITY MODELS

Criticisms have been raised about the degree to which large-scale assessments conducted at provincial and state levels can be used meaningfully for evaluating school effectiveness in accountability models (Crundwell, 2005; Ercikan, 2004, 2006a, 2006b). The major protest has been that large-scale assessments do not measure valued outcomes such as problem solving, critical thinking, and integration of knowledge, but rather provide limited information about achievement based on a single data point—the overall assessment score at the end of a school year. For the purposes of demonstrating accountability, league tables ranking schools based on large-scale assessments have been produced in Great Britain (Goldstein & Spiegelhalter, 1996; Morrison & Cowan, 1996), while state report card tables have been published in the United States from state assessment results (Coe & Fitz-Gibbon, 1998). In Canada, Cowley and Easton (2006) of the Fraser Institute have been producing annual report cards on secondary schools using provincial assessment results since 1998, and more recently they began ranking elementary schools. There has been a propensity for some researchers to be narrowly focused on the task of ranking schools rather than on establishing factors, which could explain school differences (Schafer, 2003). They have used performance data, unadjusted for intake or context, for comparing achievement across schools and reporting these in rank order in publicly accessible report cards or league tables (Cowley & Easton, 2006; Goldstein & Spielgelhalter, 1996; Morrison & Cowan, 1996). The school rankings typically have relied on single-year scores and ignored intake (e.g., prior performance and student background characteristics) and contextual effects (e.g., school processes). This resulted in inappropriate interpretations of test scores as indicators of school quality and effectiveness.

The Fraser Institute report cards have generally been directed to the parents and students as a tool to use in determining which school is the best for the student to attend rather than for government accountability (Cowley & Easton, 2006). Most ministries of education in Canada do not support the ranking of schools based on provincial assessment results. However, in B.C., where school districts enter into

an accountability contract with the Ministry of Education, the possibility of school-to-school comparisons could be made based on the set of performance indicators defined in these contracts. This has raised considerable concerns for the provincial assessment programs on the consequences of using assessments to rank schools. In B.C. over the last two years, teachers have been invited to boycott B.C. Foundation Skills Assessments by the B.C. Teachers' Federation. The main source of teachers' dissatisfaction has been their belief that the assessment results are used to judge them and their schools unfairly, and that these assessments have a negative effect on student learning.

Research has provided consistent and convincing evidence that these types of school performance reports cannot meaningfully indicate school quality or effectiveness (Aitkin & Longford, 1986; Coe & Fitz-Gibbon, 1998; Crundwell, 2005; Ercikan, 2004; Goldstein & Thomas, 1996). There has been a growing awareness that in order to investigate how effective a school is, it is necessary to determine the contribution of the school to the learning process and therefore to the learning outcomes (Gibson & Asthana, 1998; Schagen & Hutchison, 2003). This requires determining what valued outcomes are, how they should be measured, and what types of criteria or standards will be used to demonstrate progress or areas of improvement (Gray, Jesson, Goldstein, Hedger, & Rasbash, 1995; Opdenakker & Van Damme, 2001).

PURPOSES OF PROVINCIAL ASSESSMENTS OF LEARNING

In provincial assessments, students in selected grades (for example Grades 3, 6, and 9 for Alberta and Ontario, and Grades 4 and 7 in British Columbia) have been administered large-scale standardized assessments to measure learning in key school subjects such as reading, writing, and mathematics. These assessments have been designed to measure student learning outcomes based on a common set of curricular goals within the province, and to inform education policy and practice (Crundwell, 2005; Resnick, 2006; Taylor & Tubianosa, 2001). Student performance on the assessments has been used to evaluate to what extent students finishing a certain grade level are meeting the goals of learning identified by the provincial ministries of education in selected curricular areas (Crundwell, 2005; Taylor & Tubianosa, 2002). This performance information has then been used in a variety of ways by the ministries of education, school boards, schools, teachers, the public, and institutions like the Fraser Institute. These uses include informing curriculum development and planning, for accountability, and in some instances for informing

teaching and learning in the classroom. Even though provinces have commonalities in design and uses of large-scale assessments, there are some differences across provinces (Taylor & Tubianosa, 2001). The specific ways each province uses assessment data are beyond the scope of this article. Instead, the article focuses on general assessment design and development issues that have significant impact on the validity of inferences that need to be taken into account in provincial large-scale assessments.

ISSUES, DESIGN REQUIREMENTS, AND SUGGESTIONS FOR LARGE-SCALE ASSESSMENTS

This section discusses issues in current large-scale provincial assessments in providing useful information to key stakeholders in education: students, teachers, school administrators, and policy makers. Based on previous research on large-scale assessments, validity, and accountability, five basic assessment design requirements are identified and discussed. Arguments supporting each of the following five requirements are presented in the sub-sections below:

1. assessments need to assess valued outcomes;
2. assessments need to be aligned with learning, instructional, and curricular goals;
3. assessments need to provide accurate estimates of student knowledge and competencies;
4. score reports need to be informative and developed in a timely manner;
5. consequential aspects of assessment results need to be important considerations.

Assessment of Valued Outcomes

The issue of whether the assessment focuses on valued learning goals such as problem solving, critical thinking, and integration of knowledge instead of recall of factual knowledge has been the focus of much debate since the late 1980s (Linn et al., 1991; Schafer & Moody, 2004; Shepard, 2000). For about two decades, the education community at all levels, practitioners, policy makers, and researchers tackled the issue of whether certain item types were better at assessing more complex skills, such as open-ended questions versus multiple-choice questions, and whether what the items assess could be identified by their format. Some states in the United States and provinces in Canada invested in performance assessments to capture more valued learning outcomes as a response to criticisms that assessments

containing only multiple-choice had detrimental effects on education. Multiple-choice tests were criticized for narrowing down the classroom curriculum in an effort to prepare students for provincial or statewide assessments (Shepard, 2000; Shepard & Kirst, 1991).

Research focusing on cognitive demands of test items demonstrated that what test items assessed could not be determined by their format only (Baxter & Glaser, 1998; Glaser & Baxter, 2002; Leighton, 2004; Linn et al., 1991). However, there is tremendous evidence to suggest that the curricular content and skills the assessments focus on have great effect on what is taught by teachers in classrooms (Shepard, 2000; Shepard & Kirst, 1991).

Alignment With Learning, Instructional, and Curricular Goals

Research strongly suggests that external examinations not closely tied to classroom curriculum and instruction do not provide valid evidence of student learning (Linn et al., 1991; National Research Council, 2001). Furthermore, they can have substantial adverse effects on learning by narrowing the scope of classroom focus to what is being assessed by the test, resulting in emphasis on basic skills to the exclusion of promoting critical thinking and problem solving (Shepard, 2000; Shepard & Kirst, 1991).

The alignment of assessment with instruction and curricular goals is critical to supporting learning. If the aspects of learning that are assessed and emphasized in the classroom are not consistent with those targeted by large-scale assessments, students will be judged on knowledge and competencies that they may not have had opportunities to learn at school. The result will be a measure of learning occurring outside schooling contexts, where meaningful links cannot be made between schooling and the learning outcomes captured by the assessment. Therefore, the match among assessment, curricular, and instructional goals is central to the validity of inferences related to schooling and learning that can be made on the basis of the assessment.

Based on research that promotes linkages between learning and large-scale assessments (Chudowsky & Pellegrino, 2003; Ercikan, 2006b; National Research Council, 2001), some key assessment design elements are identified. The first and foremost requirement is the clarity about the underlying constructs to be assessed. One of the challenges in developing assessments that take into account the

development of knowledge and competence is the limited amount of knowledge about learning in particular content areas, and how this progression of learning can be linked to performance on assessment tasks. There are some key aspects of learning models in substantive areas that are critical to assessment design. These models need to be based on empirical studies of learners in the content domains such as reading, mathematics, and so on, rather than theoretical expectations of how students learn in a particular area. They should provide sufficient information in order to allow assessment developers to design test questions that differentiate between levels of student knowledge and skills. The student learning models also need to have detailed and specific learning goals to enable assessment developers to target them as well as to enable users of assessment information to guide learning. This will allow assessment to lend itself to being aggregated at different grain sizes so that it can be used for different assessment purposes (e.g., to provide fine-grained diagnostic information as well as coarser-grained policy-related summary information). In addition, linkages between learning models and broader learning, such as provincial curricular goals, will allow school and district administrators and policy makers to draw connections between assessment results and these learning goals.

One challenge in developing assessments that take into account the development of knowledge and competence has been limited understanding of the relationship between learning and performance on assessments (Chudowsky & Pellegrino, 2003; National Research Council, 2001). This suggests the need for close collaboration among educational psychologists and curricular and assessment experts. For example, such collaborations would involve educational psychologists providing expertise on how students learn mathematics, mathematics curriculum experts describing the appropriate content domain, and assessment experts helping to design assessments that incorporate cognitive development and content area related requirements in the assessment. Even though collaborations among curriculum and assessment experts are common, collaborations with educational psychologists who can provide the key knowledge about student learning processes are rare in large-scale assessment contexts, but necessary in order for these assessments to inform learning.

Accuracy of Estimates of Student Knowledge and Competencies

It is essential to recognize that assessment results are only estimates of the knowledge and skills of an individual. In order for data from

provincial assessments to lend themselves to valid interpretations about student learning and growth, the scores need to be reliable and generalizable. Most provincial assessments report student performance in terms of proficiency levels, such as "Does Not Meet Expectations," "Meets Expectations," and "Exceeds Expectations." Proficiency levels classify student performance into categories and are intended to be easier to interpret by teachers, parents, and students. However, research has shown that the accuracy of such classifications is critically dependent on the number of test questions as well as the quality of test questions (Ercikan, 2006c; Ercikan & Julian, 2002). This research has demonstrated that test developers should design tests and the number of test questions with their intended number of proficiency levels in mind. Classification accuracy, defined as agreement of classifications based on true and observed scores, is affected by measurement accuracy, and decreases as larger numbers of proficiency levels are considered. In their research, Ercikan and Julian (2002) found that for a given reliability level, the classification accuracy decreased on average by 0.1 points for an increase of one proficiency level, 0.2 for an increase of two proficiency levels, and 0.2 to 0.3 for an increase of three proficiency levels. In addition, classification accuracy was more sensitive to measurement accuracy when larger numbers of proficiency levels were considered. In other words, change in classification accuracy with changes in reliability was greater when higher numbers of proficiency levels were considered. Ercikan and Julian (2002) provide more detailed guidelines that can inform decisions about (a) number of proficiency levels to use in an assessment, (b) expected level of classification accuracy for an assessment with predetermined number of proficiency levels, and (c) test length for a desired level of classification accuracy and number of proficiency levels. For example, in order to determine the number of proficiency levels for an assessment, the first thing the assessment developers need to decide about is a classification accuracy level that would be acceptable for the consumers of the assessment results, such as educators and policy makers.

In addition, validity of inferences about what each proficiency level score indicates depends on the degree to which the knowledge, skills, and competencies assessed by the test are representative of the domains of knowledge and skills targeted by the provincial curricula. A test that covers all aspects of intended curricular goals in a particular content area requires administration of a large number of test questions, or tasks, in order to have a generalizable measure of achievement (Gao et al., 1994; Linn, 1993). This is often difficult

to realize due to limitations in resources and the reasonableness of testing time for students. As a result, two assessment scenarios may occur: either a short test is administered that covers a sampling of curricular goals, such as the B.C. Foundations Skills Assessments conducted by the B.C. Ministry of Education; or test items that cover the entire content domain are administered across multiple test forms of which only one is administered to each student, for example in the case of the School Achievement Indicators Program conducted by the Council of Ministers of Education Canada (Taylor & Tubianosa, 2001). Both of these scenarios may result in very small numbers of items assessing a particular learning goal. This in turn will limit accuracy of scores as indicators of knowledge and skills in relation to that particular learning goal.

The accuracy requirements for scores are different depending on the stakes associated with the decisions the scores are intended to support. For example, a short test with a reliability of 0.80 would be expected to have a classification accuracy of 0.70 approximately for four proficiency levels (e.g., Below Basic, Basic, Proficient, Advanced) (Ercikan & Julian, 2002). Therefore, on the average, approximately 30% of students may be misclassified into incorrect proficiency levels. If the scores are used as one of many sources of information about student knowledge and competencies by teachers and students themselves to guide learning, this level of inaccuracy in scores may be acceptable. However, if these proficiency level scores are used for pass/fail decisions or to assign students to remedial classes, this level of accuracy would not be sufficient, since such a high level of misclassification would have serious consequences for large numbers of students and on resource allocation. For more accurate classifications of these students, a long test with a reliability of 0.95 for four proficiency levels would be needed to have an expected classification accuracy of 0.80. Another solution would be to have a test with fewer proficiency levels, such as three, with a reliability of 0.95, which would then have an expected classification accuracy of 0.90 (Ercikan & Julian, 2002).

Score Reports Need to Be Informative and Timely

The ultimate products of large-scale assessments are the score reports. The results from assessments are communicated to different users of the assessment information through score reports for students, schools, and school districts. Therefore, in order for assessments to serve learning, the report cards and the reporting procedures need

to reflect the considerations that are necessary for assessments to inform learning. Three important aspects of the interpretation model and the reporting of scores to inform learning are discussed. These are *usefulness*, *interpretability*, and *timeliness* of score reports.

Usefulness and Interpretability of Score Reports

In order for report cards to be useful, be viewed as meaningful by students, and guide student learning, results on score reports should be targeted to specific learning outcomes. Scores that provide information about an overall performance on the assessment are of little use to students to identify which knowledge, skills, and concepts they need to target to improve their learning. The score reports should be written in a way that is free from jargon, is easy to understand, and includes instructions on how scores should be used to interpret performance. This implies that student performance on the assessment will be explained in terms of student behaviour and competencies. Similarly, in order for teachers to use score reports to guide instruction, scores must be tied to curricular and instructional goals and specific skills and competencies.

Teachers make observations of and professional judgements about student learning and competencies on a continuous basis. Therefore, for assessment results and the score reports to be treated as trustworthy, in other words as reliable indicators of student competencies, they need to be consistent with teachers' evaluations of students. Inconsistencies between teacher evaluations and provincial assessments can be indicators of the mismatch between teacher instructional and curricular goals and the provincial examinations. Such a mismatch will limit the use of assessment results to inform learning and instruction.

Timeliness of Score Reports

Typically, provincial assessments are administered at the end of an academic year to assess student learning outcomes of a particular grade. Performance results are provided to schools in score reports. In some assessments, these include individual student score reports in addition to scores at higher levels of aggregation such as class or school; in others these report cards may be reported at the classroom or school levels only (Taylor & Tubianosa, 2001). Two processes involved in creating report cards are time consuming: (a) scoring of student response papers if tests contain open-ended questions; and

(b) psychometric analyses to develop scores that have the appropriate psychometric properties, such as comparability of scores across test forms or across years. Often, the time it takes to complete these two processes results in reporting of student performance on the assessments at the beginning of the following academic year.

The score reports and therefore information about student performance on the assessment need to be provided to students, teachers, and schools in a timely fashion in order to have the greatest impact on learning. Research suggests that immediate feedback to students about their performance is one of the most effective approaches for guiding student learning (Black & Wiliam, 1998). Therefore, timeliness in reporting results to teachers and students is of critical importance for guiding learning. If teachers do not get assessment results until the following academic year, the use of assessment results is very limited. These results cannot be used to guide individual student learning since the students have moved on to another grade with different curriculum goals, and will be taught by different teachers. The teachers may use the results to evaluate learning and performance for groups of students from the previous academic year. However, this information has little value for their teaching in the current year, because the performance of one group of students in an academic year often is quite different from the performance of a group of students in the following year (Aitkin & Longford, 1986; Coe & Fitz-Gibbon, 1998; Goldstein & Thomas, 1996).

Given the critical importance for schools, teachers, and students of receiving reports of performance on assessments in a timely fashion, alternative assessment scenarios may also be considered. One alternative assessment scenario is the *embedded assessment* system. An example of such an embedded assessment is the Berkeley Evaluation and Assessment Research (BEAR) system in which student assessment is integral to instruction (Wilson & Sloane, 2000). This embedded system provides teachers with a set of tools to assess student learning and progress over time, and to provide feedback to students and other education stakeholders. The embedded nature of these assessments provides opportunities for their use in enhancing learning and the possibility of accumulating data over time that can be used for accountability purposes.

The three aspects of score reports discussed in this section cannot be addressed by large-scale assessments unless they are taken into account in the design of assessments. These aspects of score reports have

major implications for how test items are designed to assess specific knowledge and competencies, test construction, psychometric procedures, analyses, and reporting. Creating close connections between large-scale assessments and classroom instruction and learning may address both the timeliness and consistency with teacher evaluations, such as in the case of embedded assessments.

Consequential Aspects of Large-scale Assessments Need to Be an Important Consideration

The application of assessment results at the school level has been one of the most common uses of provincial large-scale assessments. In fact, this has been the typical use of assessments in accountability models to evaluate school effectiveness (Goldstein & Thomas, 1996; Taylor & Tubianosa, 2001). Research on accountability models demonstrated that the narrow focus of accountability systems on student performance, gauged by annual provincial assessments, cannot provide appropriate or sufficient information about the quality of schooling and education (Crundwell, 2005; Ercikan, 2004; Goldstein, 1997; Linn, 2000). School results on provincial assessments have been primarily based on outcome measures and neglect to include any intake measures, such as student prior attainment or background information. Performance on provincial assessments depends on many factors other than learning in schools in a particular academic year. Morrison and Cowan (1996) argued that there was a distinction between reporting the results from performance indicators in performance tables and reporting the results in rank-ordered tables. The former would more appropriately be defined as a profile that included qualitative as well as quantitative information to describe the school. These could include student characteristics—such as the knowledge and skills the students had at the beginning of the academic year and the socio-economic or family background of the students—as well as school characteristics such as socio-cultural context of schools and the schooling process. In fact, research has shown that school outcomes based solely on provincial assessment results are more highly correlated with factors external to the schooling process, such as parental education levels (Ercikan, 2004) and socioeconomic status (Thorpe, 2006), and less with internal factors, such as school context or teacher instructional practices (Ercikan, 2006a). This argument was the basis for the development of an approach known as value-added outcomes (Aitkin & Longford, 1986; Crundwell, 2005; Gibson & Asthana, 1998; Goldstein et al., 1992; Gray et al., 1995). Value-added outcomes are measures of academic

performance that provide baseline information to the researcher with which to measure the value added by schools in the learning process of students.

In general, researchers have demonstrated that there are serious and inherent limitations to the usefulness of many performance indicators for providing reliable judgements about schools and making comparisons across schools (Goldstein & Spielgelhalter, 1996; Goldstein & Thomas, 1996; Goldstein & Woodhouse, 2000; Yang & Goldstein, 1999). Overall, results from their studies have suggested that schools in these datasets vary along many dimensions and that the majority of schools in these samples cannot be statistically distinguished from each other based only on examination results. Those who develop rank-ordered or league tables should accept the responsibility of validating the variables included in their analyses, and be liable for the consequences of their use (Goldstein & Spielgelhalter, 1996; Morrison & Cowan, 1996).

Consumers of assessment performance data on schools for accountability purposes, such as ministries of education, school boards, and school districts, need to be aware of these inadequacies of school performance data based on single-point assessments as indicators of quality of education. The inadequacies summarized above strongly suggest that the interpretation of single-point assessments has poor validity and negative consequences for school accountability.

SUMMARY

Provincial large-scale assessments have played a very important role as the primary data collection effort about school learning outcomes. As discussed throughout this article, the validity of interpretation of assessment data as an indicator of school learning has been based on the meaningfulness of performance, the quality of the data, and how such data has been used and interpreted. Based on these three aspects of large-scale assessments, this article focused on identifying issues with design and validity and providing recommendations for large-scale assessments to better serve the multiple purposes they are intended to serve. We argued that meaningfulness of performance on assessments is critically dependent on two factors: whether the assessments focus on valued outcomes and the degree to which they are aligned with learning, curricula, and instruction. These factors in turn have many implications for what types of test items are used; what content, knowledge, and skills the assessment focuses on;

and whether these correspond with quality education. The quality of data from assessments is dependent on the accuracy of the scores obtained from the assessment, which are determined with respect to the decisions the assessment data are intended to inform. If assessment results are used for high-stakes decisions, such as graduation requirements, they will need to have sufficient number of items to result in scores with an acceptable level of classification accuracy. If, on the other hand, the assessment results are intended to be used for informing learning and instruction, they need to provide accurate scores on targeted learning outcomes.

Use and meaningfulness of interpretations of scores are tied to how the scores are reported. Many aspects of score reports can contribute to use and meaningfulness. The types of scores reported, a verbal description of performance on assessments, and guidance to users of assessment data about how to interpret test results and cautionary statements about limitations of interpretations will all contribute to use and meaningfulness.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society Series A (Statistics in Society), 149*, 1–43.

Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 37–45.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*, 1–57.

British Columbia Ministry of Education. (2007). *Graduation program requirements*. Retrieved October 12, 2007, from <http://www.bced.gov.bc.ca/graduation>.

Carnoy, M. (2005). Have state accountability and high-stakes tests influenced student progression rates in high school? *Educational Measurement: Issues and Practice*, *24*(4), 19–31.

Chudowsky, N., & Pellegrino, J.W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*, 75–83.

Coe, R., & Fitz-Gibbon, C.T. (1998). School effectiveness research: Criticisms and recommendations. *Oxford Review of Education, 24*, 421–438.

Cowley, P., & Easton, S. (2006, April). *Report card on secondary schools in British Columbia and Yukon: 2006 edition*. Vancouver, BC: Fraser Institute.

Crundwell, R.M. (2005). Alternative strategies for large scale student assessment in Canada: Is value-added assessment one possible answer? *Canadian Journal of Educational Administration and Policy*, *41*, 1–21.

Earl, L.M. (1999). Assessment and accountability in education: Improvement or surveillance. *Education Canada*, *39*, 4–6.

Ercikan, K. (2004, March). *Limits and constraints on interpretations from large-scale assessments*. Paper presented at the UBC Assessment Conference, Vancouver, BC.

Ercikan, K. (2006a, October). *Assessment, accountability, and equity*. Paper presented at the British Columbia Teachers Federation Conference: Counting What Counts, Vancouver, BC.

Ercikan, K. (2006b). Developments in assessment of student learning and achievement. In P.A. Alexander & P.H. Winne (Eds.), *American Psychological Association, Division 15, Handbook of educational psychology* (2nd ed., pp. 929–952). Mahwah, NJ: Erlbaum.

Ercikan, K. (2006c). Examining guidelines for developing accurate proficiency level scores. *Canadian Journal of Education*, *29*, 823–838.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15,* 269–294.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, *7*, 323–342.

Gibson, A., & Asthana, S. (1998). Schools, pupils and examination results: contextualising school performance, *British Educational Research Journal, 24*(3), 255–268.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355–392.

Glaser, R., & Baxter, G.P. (2002). Cognition and construct validity: Evidence for the nature of cognitive performance in assessment situations. In H. Braun, D. Jackson, & D. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 179–191). Mahwah, NJ: Erlbaum.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369–395.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1992). Multilevel models for comparing schools. *Multilevel Modelling Newsletter, 4*(2), 5–6.

Goldstein, H., & Spiegelhalter, D. 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of Royal Statistical Association, 159*, 385– 443.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society Series A (Statistics in Society), 159*(1), 149–163.

Goldstein, H., & Woodhouse, G. (2000). School effectiveness and educational policy. *Oxford Review of Education, 26*(3/4), 253– 363.

Gray, J., Jesson, D., Goldstein, H., Hedger, K., & Rasbash, J. (1995). A multi-level analysis of school improvement: Changes in schools' performance over time. *School Effectiveness and School Improvement, 6*(2), 97–114.

Leighton, J.P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1–16.

Linn, R.L. (2000). Assessment and accountability. *Educational Researcher, 29*, 4–16.

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

Mislevy, R., Wilson, M., Ercikan, K., & Chudowsky, N. (2002). Psychometric principles in student evaluation. In D. Nevo & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 478–520). Dordrecht, Netherlands: Kluwer.

Morrison, H.G., & Cowan, P.C. (1996). The state schools book: A critique of a league table. *British Educational Research Journal, 22*, 241–249.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*, Washington, DC: National Academies Press.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575–603.

Opdenakker, M-C., & Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effects on mathematics achievement. *British Educational Research Journal, 27*, 407–432.

Resnick, L. (2006, Spring). Making accountability really count. *Educational Measurement: Issues and Practice, 25*(1), 33–37.

Schafer, W. (2003). A state perspective on multiple measures in school accountability. *Educational Measurement: Issues and Practice, 22*(2), 27–31.

Schafer, W.D., & Moody, M. (2004). Designing accountability assessments for teaching. *Practical Assessment, Research & Evaluation, 9*, 14. Retrieved February 22, 2007, from <http://PAREonline.net/gevn.asp?v=9&n=14>.

Schagen, I., & Hutchinson, D. (2003) Adding value to educational research: The marriage of data and analytical power. *British Educational Research Journal*, *29*, 749–765.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4–14.

Shepard, L., & Kirst, M.W. (1991). Interview on assessment issues with Lorie Shepard. *Educational Researcher, 20*, 21–25.

Sicoly, F. (2002). Stability of school-level scores from large-scale student assessments. *Applied Measurement in Education, 15*, 173–185.

Stone, J.E. (1999). Value added assessment: An accountability revolution. In M. Kanstoroom and C.E. Finn, Jr. (Eds.), *Better teachers, better schools* (pp. 239–250). Washington, DC: Thomas B. Fordham Foundation.

Taylor, A.R., & Tubianosa, T. (2001). *Student assessment in Canada: Improving the learning environment through effective evaluation*. Kelowna, BC: Society for the Advancement of Excellence in Education.

Thorpe, G. (2006). Multilevel analysis of PISA 2000 reading results for the United Kingdom using pupil scale variables. *School Effectiveness and School Improvement, 17,* 33–62.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181–208.

Yang, M., & Goldstein, H. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education, 25*, 469–483.

Kadriye Ercikan, Ph.D., is an associate professor in the Faculty of Education at the University of British Columbia. She specializes in measurement, evaluation, and research methodology. Her research has most recently focused on validity and fairness issues in multicultural and multilingual assessments, construction of data using multiple methods for research, and psychometric issues in large-scale assessments.

Stephanie Barclay McKeown, M.A., is a Ph.D. student in the Faculty of Education at the University of British Columbia, specializing in measurement, evaluation, and research methodology. Her research interests include validity issues with assessment design and analysis, evaluation, large-scale assessments, and program and policy planning.