Alternative Strategies for Large Scale Student Assessment in Canada: Is Value-Added Assessment One Possible Answer

R. Marc Crundwell, University of Michigan, Dearborn

Abstract

Recent focus on student achievement and the effectiveness of schools, school boards, and teachers has lead to increased demands for accountability in education. Large scale assessments are now used in most provinces in Canada to examine the degree to which educational standards are being reached and explore issues of accountability. Alternative models of accountability such as value-added models are gaining popularity in other countries. The current paper explores weaknesses of large-scale annual assessment and investigates the degree to which value-added models may be helpful in looking at educational accountability.

Introduction

The publication of the 1983 National Commission of Excellence in Education document, *A Nation at Risk*, brought to the forefront the issue of educational accountability in the United States (U.S.). The recommendations presented by the commission almost single handily initiated the call for educational accountability on the part of federal and state governments in education. In fact, the commission report played a key role in the development of large scale high stakes assessment procedures in the U.S. designed to address educational accountability and rate how schools were performing (Barksdale-Ladd & Thomas, 2000; Michigan Association of School Psychologists [MASP], 2004). In the move towards accountability state governments have mandated public reporting of results so that the general public would have information on the performance of their local schools. As a result of these reports, the assessments are well known to have an effect on the value of homes in an area, the ability to draw industry to that area, parent involvement in the school, as well as the ability to pass bonds or other funding issues relating to a school or a district (MASP, 2004). For some parents, selection of a community in which to live is based fully on the rankings of the schools within an area.

More recently, the passing of the Elementary and Secondary Educational Act (ESEA) has placed increased emphasis on the role of the U.S. federal government in education. ESEA, better known as No Child Left Behind (NCLB), places increased emphasis on standards-based accountability and mandates large scale assessment from grade 3 to grade 8 to ensure continued progress towards academic proficiency and standards (MASP, 2004). Within the NCLB legislation, the performance of a school or district on the assessment can result in either incentives or severe sanctions for schools that meet or fail to meet predetermined levels of proficiency. Schools that fail to meet the predetermined threshold of proficiency may lose funding or be taken over by the federal government. NCLB legislation has again increased controversy over the use and interpretation of large scale assessments in the U.S.

Presently, four issues that parallel those of the U.S. can be observed in Canada. First, public education in Canada is under increased scrutiny and there is increased pressure for education

systems and teachers to show that what is happening in the schools in working (Earl, 1999). For many individuals there has been a steady growth of concerns regarding the educational accountability of schools and teachers with regard to student achievement. The consensus of many Canadian's is that the educational system in Canada is no longer providing the level of education needed in today's society and that schools are failing (Earl, 1999; Hepburn, 1999). Based on these widely held beliefs, it is not surprising that concerns about accountability have become a key focus of politicians and political parties within the provinces. As a result, both policymakers and parents have increased their attention on the quality of public schools and are calling for improved standards and accountability. While public and political calls for accountability in education are typically accepted, the ability to define and measure it is exceedingly more complex (Wickstrom, 1999).

In response to concerns about the quality of public education, the second issue that has developed in Canada involves an increased demand for object measures and assessment tools that will provide evidence that students and schools are meeting achievement standards (Stone, 1999). Miles and Lee (2002) have identified the recent public agenda to improve education through mass testing programs and have defined this concept as political validity. In this regard, large scale assessments have become the vehicle of choice in the U.S. and in Canada to address achievement and accountability. This focus has increased the use of large scale tests to make decisions about students and to hold educators and schools accountable for outcomes (Earl, 1999). Current research has shown that public support for large scale assessments is both consistent and longstanding (Phelps, 2005). At the current time, nine of the ten provinces in Canada conduct some form of large scale assessment to examine the academic achievement of students. Most provinces also attempt to gage the effectiveness of schools based on the results of these assessments, although at least one province does not disaggregate the data of these assessments across school boards or individual schools. In Canada, as in the U.S., there is no universal or uniform large scale assessment system, with each province having developed their own assessment tools, administration procedures, and selected grades of administration. In general, current assessments in elementary schools are used to determine achievement levels of students. In high schools large scale assessments are used for a variety of purposes in different provinces, from assessing skill development, assessing prerequisite skills and combining with

work completed in the school to comprise a final grade, to those that must be passed to allow a student to graduate.

As the use of large scale assessments has grown in Canada, a third concern has been the communication of test scores to the general public. This includes reporting about individual student achievement, as well as about achievement for an individual school or school board. As Earl (1999) has noted, large scale assessments suggest to the general public that the effectiveness of students, schools, and teachers can be moved into the world of statistics in which there is increased accuracy and objectivity in examining achievement. However, as is often the case, statistics can be misused and misinterpreted. Simner (2000) has discussed the dangers of how large scale assessment data is reported to the general public and the inappropriate conclusions that are drawn from the method of reporting, especially when released in averages. When presented in simple form, many people fail to realize that a single score can not adequately sum up a complex construct such as achievement (Earl, 1999).

The final issue that parallels those experienced in the U.S. involves the psychometric properties of the assessments, as well as the use of data from these assessments to draw conclusions regarding the accountability of schools. Large scale assessments used in Canada often lack strong reliability and validity data and are often developed without consideration of accepted test development standards to assure psychometric soundness (Miles & Lee, 2002). Very few provinces currently provide any documentation to the public regarding the reliability and validity of their assessments, or report having data on these issues that is only available internally. While lacking this data, such assessments are still used to make decisions about individual students and the effectiveness of schools and teachers with regard to the achievement of children. As Messick (1989) has noted, we also need to be concerned about how the results of test scores are interpreted. Messick (1989) refers to this concept as consequential validity and stresses that we must be sure that the use of test scores is justified based on the properties of the test and the testing situation. Further, if the goal of large scale assessments is to improve education, then the assessment instruments must be psychometrically sounds (Miles & Lee, 2002). Related to the above issue is concerns that many accountability systems are build only on large scale

assessments, and that the results of these are then used to influence and make education policy and curriculum decisions (Cizek, 2005).

What Is Educational Accountability?

Quite surprisingly, accountability in education is difficult to accurately define. In general, accountability refers to an individual's or an organization's responsibility for developing and implementing a process or procedures to justify decisions made and to demonstrate the result or outcomes produced (Council of Chief State School Officers, 2005). Within the field of education, Adams and Kirst (1999) have indicated that views of what educational accountability is have evolved and changed as public schools have evolved and changed, and have typically followed economic and political movements. While Adams and Kirst (1999) have described six different types or models of accountability, which include bureaucratic, legal, professional, political, moral, and market accountability, they are rarely considered or discussed by most individuals. In contrast, the form of educational accountability that is most pervasive in the minds of the majority of individuals and draws the most attention and conversation, is based on the societal belief that schools are responsible for demonstrating their contributions to student learning. This form of educational accountability holds students, schools, and school boards responsible for academic achievement (Elmore, 2002). In fact, the push for accountability of academic achievement for all students is at a level that is unprecedented in comparison to previous decades or generations (Sanders, 2000).

Large Scale Assessment in Canada

Large scale assessment of students is typically viewed as an important component of determining achievement in education and should be carried out through a systematic fashion (Taylor & Tubianosa, 2001). While such assessments are a complex undertaking, they are felt to provide data that is crucial in making decisions regarding education. According to Taylor and Tubianosa (2001), large scale assessment and evaluation in education is important for a number of reasons that include identification of individual strengths and weaknesses. With regard to the current discussion, the last two functions of large scale assessed discussed by Taylor and Tubianosa

(2001) areas are key. First, large scale assessment can provide an evaluation of the effectiveness of educational programs implemented in schools, as well as provides information to the general public on how schools and students are performing and how they compare with other schools. Performance on these assessments can also be linked to accountability or funding. Further, such data allows for the development of interventions to improve student achievement and quality of the educational system. Second, large scale assessment practices can also be used as selection mechanisms that are purported to be fair to all individuals taking them. In this regard they may be used for grade promotion or graduation determination.

As previously discussed, there is great variability in the development and administration of large scale assessments across provinces. In general, all provinces except Prince Edward Island have some form of large scale assessment for both elementary age and for high school age students. These assessments are normally used to determine the achievement of individual students, as well as to gather performance data about individual schools and boards of education within the province. A review of each province's assessment practices is beyond the scope of this paper. While great variance does exist between the individual provinces's, issues relating to the psychometric properties of these assessments and the use of data to draw conclusions about the state of education in each province are similar. Ontario is one of the few provinces to provide some reliability and validity data regarding their assessments. As a result the remainder of this article will examine the Ontario Quality and Accountability Office (EQAO) assessment that is administered in Ontario. It is believed that in general the issues discussed in this paper can also be applied to the assessment practice of other provinces.

In Ontario, large scale assessments with sections assessing Reading, Writing, and Mathematics are administered annually to students in grades 3 and 6 at the elementary level. At the high school level, large scale assessment in Mathematics is completed annually in grade nine. A literacy test is administered in grade 10 and students require a passing mark to graduate from high school. Every year each school board in the province must publish the results of these assessments. According to the EQAO (EQAO, 1998a; 1998b), the assessment and publication of the results provide a base of information that allows the public the ability to make judgments about the quality of education available across Ontario. EQAO suggests that educational

accountability is achieved through the large scale assessment in numerous ways. In terms of taxpayers, the assessment provides information on how students within the system are achieving. When compared against accepted and understandable standards, it allows an evaluation of student success and ways to improve learning. The assessment also allows the general public the ability to compare student achievement in Ontario to national and international standards. Finally, as the results are made known, the general public is able to determine the state of Ontario's schools with the supplied data on achievement, as well as have an idea of the contextual factors that influence student learning.

Weaknesses of Large Scale Assessments

Student achievement and accountability in education are important areas that must be addressed through on-going assessment and evaluation. Large scale assessments used in the majority of Canadian provinces can be an important and helpful component in understanding student achievement and in analyzing and evaluating the effectiveness of the educational system. While such assessments can play a key role, a number of inherent weaknesses in the current assessment practices reduce the ability to draw conclusions regarding accountability as expected by the general public. These weaknesses will be discussed below with a focus on the assessment techniques used in the province of Ontario.

In Ontario, the current assessment materials and practices suffer from a number of issues. First, there are issues related to the reliability of the assessment instrument. Reliability refers to the consistency of measurement of an assessment tool and provides an estimate of the measurement error within a given score (Anastasi & Urbina, 1997). Reliability can be examined in a variety of ways including test-retest measures of score consistency over time, internal consistency measures that explore the homogeneity of test items, alternate form reliability which examines consistency across different forms of a test, and inter-rater reliability that represents the consistency in scoring the assessment between different individuals.

Wolfe, Wiley, and Traub (1999) explored inter-rater reliability in a paper written under contract with EQAO. In their paper, Wolfe, Wiley and Traub (1999) used generalizing theory to explore

the issue of reliability, with a focus on inter-rater reliability. Generalizing theory is reported to be an expansion and extension of the methods used in traditional reliability theory. It considers measurement reliability from samples to relevant populations and takes into account variance across the performance of different individuals, as well as variation within an individual's performance. Using their model, they report a 70% to 80% probability that a student's performance would be marked correctly. In essence, their results indicate the accuracy of the mark that was assigned to each paper in terms of the probability that it was correctly classified. They suggest that inter-rater reliability is reasonably high and comparable to other large-scale performance assessments.

Wolfe, Childs, and Elgie (2004) have also explored score reliability in a review of EQAO assessment materials and procedures. They explored the impact of the number of items on the tests to reliability, indicating that the assessment programs biggest difficulty is the number of items. Review of the data from the 2003 administration by Wolfe, Childs, and Elgie (2004) indicated that overall reliability was high for large area scores with many items. However, reliability dropped below levels generally considered necessary for reporting individual scores on sub scores that contained half as many items.

The Wolfe, Wiley, and Traub (1999) and Wolfe, Childs, and Elgie (2004) studies represent the only published reliability results for the Ontario assessment that have been released to the general public. While these studies offer important data regarding inter-rater reliability and the impact of item numbers on reliability, other important reliability data such as test-retest reliability has not been provided. Test-retest reliability is essential to evaluate a test's reliability. Also of concern is the lack of data regarding sampling error and measurement error. In the case of large scale assessments, sampling error reflects the reliability of testing a different group of students each year, while measurement error refers to the variation in scores associated with testing students on a particular occasion (MASP, 2004). Fluctuating scores on large scale assessments are typically the result of sampling error. Finally, no reliability estimates are provided for various subgroups of the population, hence no data is provided regarding consistency of the assessment for these subgroups (MASP, 2004).

The second area of weakness is the validity of the assessments used. According to Messick (1995, p. 741), "validity is an overall judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment." While conventional views of validity focus on three types of validity: content validity, criterion-referenced validity, and construct validity, some authors have questioned this view. Messick (1995) has suggested the conventional view fails to take into account other important aspects of scores. He indicates six forms of validity should be considered: content, substantive, structural, generalizability, external, and consequential. Of these, structural validity which refers to the fidelity of the scoring structure, generalizability refers to the ability to generalize results to the population and across populations, and consequential validity involves the social consequences of the assessment to the society. Each of these is of great importance in examining the validity of large scale assessments.

At this time there is no published validity data for the assessments used in Ontario. As the content of the assessment is drawn from provincial curriculum, aligned to standards, and undergoes extensive content reviews it is likely that the assessments has adequate content validity. However, as noted above, no other measures of validity have been provided. Wolfe, Childs, and Elgie (2004) indicate in their review of the assessment instrument and procedures that an active program of validity research be initiated and supported. Specifically, they indicate such research is needed to ensure appropriate and valid interpretation and use of scores from the assessment.

When trying to interpret assessment results, consideration must be given to non-instructional variables that negatively impact on such achievement tests (Grissmer, Flanagan, Kawat, & Williamson, 2000). While acknowledging non-instructional variables, current large scale assessments like those used in Ontario, do little to address these in analyzing or reporting the results. Variables such as family stability and mobility, parental involvement and expectations regarding school success, parent income level, parent education level, family ethnicity, student motivation, student absenteeism, and student capacity for learning have all been shown to influence test performance (Simner, 2000; MASP, 2004). Simner (2000) has indicated that

because students are not randomly assigned to schools and schools have little control over the majority of the factors listed above, blame for poor performance cannot be placed solely on the school. Proper consideration must also be given to each of the factors listed above.

These issues are important as studies have supported the impact of these variables on large scale assessment. Burns, Courtad, Hoffman, and Folger (2004) examined the relationship between the percentage of students receiving free or reduced breakfast and/or lunch, pupil/teacher ratio, and foundation allowance grant per pupil to results of the Michigan Education Assessment Program (MEAP) in mathematics, reading and writing in each Michigan district. Results of the study indicated that the only significant predictor of MEAP achievement scores was the percentage of students receiving free or reduced breakfast and/or lunch. The authors concluded that MEAP achievement scores were significantly influenced by community socio-economic status, a non-instructional variable. Even if the report is provided as the percentage of students at one of four levels, as in Ontario, studies have indicated that the differences between schools and school boards can be strongly related to socio-economic factors within the population of that school or school board (Sanders, 2000).

In Ontario this issue has started to be addressed by developing the Education Quality Indicators Framework. This framework reports on a range of environmental factors that impact on achievement. As student achievement results are considered key indicators of educational quality, EQAO (2004) indicates that student scores on large-scale assessments only be interpreted meaningfully in the context of the system that produced them. Further, they indicate that to understand and evaluate the quality of education, more then numerical values of achievement, but also a more comprehensive picture of the unique and complex characters of schools, boards and the province. Unfortunately, these indicators are summarized and provided only to school boards and not provided or released to the general public to assist them in understanding the scores of a particular school or area. While starting to address this issue, the impact of non-instructional variables is rarely addressed in large scale assessments.

The final weakness is that of drawing conclusions from a single data point and the method of reporting the data to the public. Large scale assessments such as those used on Ontario are often

used to make educational decisions regarding students, individual schools, as well as school boards. In these cases, decisions are made on a single test score. Even if one removes the previously noted issues of reliability and validity of large scale assessment measures, the use of a single test score to make far ranging decisions and judgments about the status of education in a province are inappropriate. Further, in the case of Ontario, the use of criterion-referenced measures does not provide a valid manner in which to compare individuals, schools, or school boards as the data is not standardized. As Simner (2000) notes in the position paper of the Canadian Psychological Association (CPA) and the Canadian Association of School Psychologists (CASP), a single assessment does not provide enough information to make accurate or responsible comparisons. Volante (2004) also notes that the presentation of large scale assessment results pits schools and districts against one another and often leads to schools adopting maladaptive test preparation (i.e., teaching to the test) that negatively impacts on student learning. In this regard, EQAO (1998a) indicates in its parents guide to the assessment, the primary purposes of the assessment in Ontario are to improve student learning, identify areas of strengths, and address areas where improvement is needed. It is not about passing or failing, or about comparing schools.

Reporting the assessment results is also an area of great concern for most large scale assessments. According to Sanders (2000, p. 332), "the worst possible use of test data for public reporting is the presentation of simple test averages by district and schools." In most cases, this is the manner in which data is presented to the public, especially by local media sources, and encourages inappropriate comparisons between schools and districts. The position papers of both the MASP (2004) and CPA/CASP (Simner, 2000) recommend results of such large scale academic assessments should not be used to rank schools or school boards. They recommend agencies responsible for assessments, work to ensure that information distributed about the appropriate use of data from large scale assessments include how results can be misused. This is consistent with the EQAO (1998b), who indicate in the educators guide to the assessment that results should not be used to rank schools or boards as ranking does not contribute to the well-being of Ontario students and is inconsistent with the EQAO's mandate and core values. Yet, the Ministry of Education in Ontario mandates the release of data in a manner that creates such comparisons.

One Possibility: Value-Added Assessment

The current focus on accountability in Canada has resulted in the development and implementation of some form of large scale assessment in the majority of provinces to examine the achievement of students, as well as the effectives of schools and school boards. In most cases, accountability has been addressed through the administration of a yearly assessment of academic achievement across selected grade levels. While in the U.S. these tests often have high stake consequences for school and school boards, those used in Canada are typically used to assess state of education in the province. In this regard, they are a snap shot of how individual students are achieving, as well as how schools and school boards are succeeding in the education of children. While they are able to provide information regarding trends in achievement, the results of these assessments are often been used to more strongly assess the performance of schools and school boards in an increasingly open manner (Raham, 1999). In general, these assessments have become a vehicle for accountability and are used to make schools accountable and to drive educational reform and policy in many provinces.

The use of a single data point to draw conclusions regarding the state and effectiveness of education within a school board or an individual school should be seriously questioned. Such data does not allow one to examine the degree to which an individual school or a school board is impacting on the education of the students under their charge. It also does not address issues related to non-instructional variables or gains made by individual students within a class or school. As large scale assessments are increasingly used to hold educators and schools accountable, a method of examining the impact of teachers and schools is required to address accountability questions (Kupermintz, 2003). One possible method of addressing these concerns and providing more useful data regarding the effectiveness individual students, schools and school boards is value-added assessment. Value-added assessment practices address this issue by measuring the educational value-added contribution by teachers and schools to student learning. Value-added assessment address the issue of annual large scale assessments and their inability to adequately determine the impact of schools in student progress by examining student

progress over time and isolating the aspects of student learning and achievement that can be attributed to educators and schools (Sanders, 2000).

The value-added assessment model is most often associated with the Tennessee Value Added Assessment System (TVAAS) developed in the late 1980's by Dr. William L. Sanders, a professor at the University of Tennessee. The model is based on Sanders (2000) belief that student learning must be viewed as a learning ramp. While the desire for each child is to move up the same ramp, it is important to recognize that not all students will move up the ramp at the same speed or interval. Further, this approach recognized that not all students reach a designated level of achievement at a specific time or at the same time. If assessment and measurement is viewed in this manner, then a model of accountability based on the progress of individual students can more accurately assess the impact of schools on academic achievement and growth. Specifically, the model allows for an examination of the instructional variables schools have control over, while controlling for those non-instructional variables in which they have no control. Using this approach, the rates of academic progress of students within a school can be estimated nearly free of socio-economic confounding variables. It is important to note that value-added assessment does not replace standards-based assessment. It offers a means to accurately determine the role of schools, school boards, and teachers in the attainment of academic skills based on student progress, rather then the percentage of students to meet an absolute standard (Stone, 1999; Ballou, Sanders, & Wright, 2004).

Value-assessment requires two key components. First, it requires the use of achievement test data that meet three conditions (Sanders, 2000): (i) the achievement test scores must correlate with curriculum objectives; (ii) the tests used must have sufficient ability to measure progress of both previously low and high scoring students; and (iii) the tests must have reliability of measurement. In the value-added model, either traditional norm-referenced or criterion-reference tests can be used as long as they correlate highly with curriculum objectives and meet the remaining two conditions. The second key component is multiple data collection points. Longitudinal data is required for each student in this model; hence, each student will need to have been assessed on a number of occasions to build a data set. The model is based on statistical mixed model theory and methodology and assesses each student's academic

achievement pathway and when aggregated over many students, provides an assessment of the school and districts effectiveness.

The Fraser Institute has reported on the effectiveness of schools in some provinces and has provided numerous documents that rank schools. In the Report Card on Elementary Education in Ontario 2004 Edition (Crowley & Easton, 2004), the authors use a value-added system to examine the role of schools in academic achievement and progress. While this system utilizes many of the aspects of a value-added system and attempts to deal with socio-economic and non-instructional variables, there are three significant issues which impact on any conclusions that can be drawn using a value-added model as described above. First, data used to determine a schools academic improvement or decline was constructed by comparing a schools result for the current year with the results from previous years. While this provides an indication of the difference between the results of each year's assessment in that school, the groups are composed of different students and the characteristics of each group would be different. A value-added model requires the use of multiple data points for each student so that each student actually serves as their own control and comparing different subpopulation of students is avoided (Sanders, 2000). Using different subpopulations reduces both conclusions that can be drawn from the data and the identification of effectiveness schools as suggested by Crowley and Easton (2004). Most provincial assessments and reports also compare different groups of students when examining results and use data from the current and previous years (i.e., different students) to draw conclusions about student achievement and effectiveness of schools. In the value-added model proposed by Sanders (2000), using individual student results of achievement overtime in comparison to achievement results of different subpopulations is reported to address issues of socio-economic and ethnic variance.

The final issue noted in the value-added approach used in the Fraser Institute report cards is that of missing data. A regression analysis requires complete data for each case (Sanders, 2000). Visual inspection of the provided data for each school indicates that a large number of schools lacked data regarding parental level of education. This variable was reported to be the key indicator used in the analysis for socio-economic status. The report does not indicate how missing data in this area was handled in the analysis.

A true value-added model addresses the issues noted above in the Fraser Institute reports regarding accountability in education, as well issues of accountability in the use of annual large scale assessments to determine the effectives of schools. Value-added assessment offers several key advantages over other forms of annual large scale assessment for educational accountability. According to Stone (1999), these advantages include the ability to examine teacher, school, and school system effectiveness by looking at the increases in achievement of individual students across a period of several years. In contrast, most current systems of accountability using annual large scale assessments assess school effectiveness by comparing current student achievement to an average or to an arbitrary standard. Value-added models also reduce the influence of pre-existing differences among students and level the playing field for teachers, schools, and systems. This includes issues such as race, socioeconomic status, previous learning, intelligence, and other non-instructional variables. Value-added assessment using mixed-models also more appropriately manages missing test data and incomplete data sets. Sanders (2000) also indicate that value-added assessment utilizes up to five years of data for each student and reduces the reliance on a single test. Hence, instead of scores for three subject areas on a grade three or grade six test, value-added assessment would examine the three subject areas and each subject areas items, resulting in a student data file that can contain hundred of data points. The large number of data points spanning over years and academic subjects provides far more information then can be provided by any one test for a specific year. Finally, the approach is not used to rank schools in any manner, but to provide a more appropriate way of examining accountability.

Overall, the value-added approach focuses on student progress over time and those in support of such a model argue that it provides a more accurate and trustworthy quantitative measure of student learning. Further, the value-added model is reported to provide results of student achievement that can be directly attributed to schools and teachers (Sanders, 2000). In doing so, the model takes care of issues related to assessment and measurement in large scale assessments using only annual assessment data.

Conclusion

Many provinces in Canada have moved towards the use large scale annual assessments to examine student achievement. While these assessments in Canada do not have high stakes consequences attached to their results like in the U.S., they have increasingly been used as a means to hold students, schools, and school boards accountable for academic achievement. While yearly assessments may provide some relevant data, they are limited in their ability to determine the effectiveness of teachers, schools, and school boards in their current form. Despite this fact, they continue to be used as a major foundation of educational accountability by politicians, think tanks, and commissions to decry the state of education. Further, they have been used to influence and direct educational decisions and policies. The current form in which most provinces release the results continue to lead to misinterpretation by many individuals and organizations, and they are often used to compare and rank schools even those they were not designed for these purposes. The Fraser Institute, for example, has released report cards for many provinces using the data from provincial assessments, as well as a simply approach to value-added procedures. However, there are significant issues with the methodology used in their analyses that violate current value-added models.

If large scale assessments will continue to be used to imply the effectiveness of schools and school boards and address educational accountability, then value-added assessment methods, like those developed by Sanders (2000) provide one option to more accurately assess accountability in education. This is especially important as the current direction in many provinces appears to lean towards the use of these assessments to suggest the accountability of schools and school boards, as well as to direct education policy and decisions. Value-added methods can provide data that can be used to address issues of curriculum, instructional strategies, and educational programs as a whole, as well as to more accurately address issues of accountability if large scale assessment results in Canada are increasingly to be used for this purpose.

While value-added approaches are one method that can be used to assist in examining accountability, it is also vital that researchers explore issues related to such systems of accountability and other methods. Researchers can provide information about the strengths and weaknesses of different accountability approaches. As Linn (2003, p. 3) indicates, “I doubt that

anyone would say that we already know all we need to know to design a highly effective accountability system that is sure to contribute to the broad goal of improving education without having major unintended negative side effects". It is likely that accountability systems should go beyond just external assessment measures and also include other dimensions that can capture issues of accountability in a systematic way (Raham, 1999). Finally, we must view educational accountability as a shared responsibility. Administrators and policy-makers must also act in a responsible manner and provide both instructional resources and professional development to allow students and teachers to meet the expectations of accountability systems (Linn, 2003). The latter group must be willing to provide the necessary financial resources as well as use the expertise of researchers to develop appropriate systems of accountability.

References

Adams, J.E. & Kirst, M.W. (1999). New demands and concepts for educational accountability: Striving for results in an era of excellence. In J. Murphy and K.S. Louis (Eds.), *Handbook of research on education administration* (2nd ed., pp. 463-489). San Francisco: Jossey-Bass.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.) Upper Saddle River, NJ: Prentice-Hall.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education*, 51(5), 384-397.

Ballou, D. (2004). Rejoinder. *Journal of Educational and Behavioral Statistics*, 29(1), 131-134.

Burns, M. K., Courtad, C. A., Hoffman, H., & Folger, W. (2004). A comparison of district-level variables and state accountability test results for public elementary and middle schools. *Psychology and Education*, 41(2), 17-26.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-65.

Cizek, G. C. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 23-54). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Council of Chief State School Officers (2005). The words we use: A glossary of terms for early childhood education standards and assessment. Retrieved November 20, 2004, from http://www.ccsso.org/Projects/scass/projects/early_childhood_education_assessment_consortium/publications_and_products/2919.cfm

Crowley, P. & Easton, S. T. (2004). *Report card on Ontario's elementary schools: 2004 edition*. Retrieved October 25, 2005, from The Fraser Institute Web site: http://www.fraserinstitute.ca/shared/readmore.asp?sNav=pb&id=665

Earl, L. M. (1999). Assessment and accountability in education: Improvement or surveillance. *Education Canada*, 39(3), 4-6,47.

Elmore, R. F. (2002). *Bridging the gap between standards and achievement: The imperative for professional development in education*. Retrieved November 15, 2004, from the Albert

Shanker Institute Web site: http://www.shankerinstitute.org/Downloads/Bridging_Gap.pdf

Education Quality and Accountability Office (1998a). *Parents Handbook*. Toronto, Ontario: Queen's Printer of Ontario.

Education Quality and Accountability Office (1998b). *Educators Handbook*. Toronto, Ontario: Queen's Printer of Ontario.

Education Quality and Accountability Office (2004). *Completing the picture: The education quality indicators framework*. Retrieved October 25, 2005, from the EQAO Web site: http://www.eqao.com/11e/11_4e.aspx?Lang=E

Grissmer, D. W., Flanagan, A., Kawat, J., & Williamson, S. (2000). Improving student achievement: What NAEP state scores tell us. Retrieved November 1, 2004, from the Rand Web site: http://www.rand.org/publications/MR/MR924/

Hepburn, C. R. (1999). The case for school choice: Models from the United States, New Zealand, Denmark, and Sweden. Critical Issues Bulletin, September 1999 Issue. Retrieved May 10, 2005, from the Fraser Institute Web site: http://oldfraser.lexi.net/publications/critical_issues/1999/school_choice/

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis*, 25(3), 287-298.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 332(7), 3-13.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan Publishing Company.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

Michigan Association of School Psychologists (2004). *Position statement on the use of the Michigan Educational Assessment Program (MEAP) for high stakes decisions for students in Michigan*. Retrieved January 2, 2005 from the MASP Web site: http://www.masponline.org/High_Stakes_Testing_in_Michigan.htm

Miles, C. A., & Lee, C. (2002, April). In search of soundness in teacher testing: Beyond political validity. Paper presented at the meeting of the American Educational Research Association, New Orleans, Louisiana.

National Commission of Excellence in Education (1983). A nation at risk: The imperative for educational reform. Retrieved November 11, 2004, from http://www.ed.gov/pubs/NatAtRisk/risk.html

Phelps, R. P. (2005). Persistently positive: Forty years of public opinion on standardized testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 1-22). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Raham, H. (1999). Linking assessment and school success. *Education Canada*, 39(3), 23, 46.

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, 14(4), 329-339.

Simner, M. L. (2000). *A joint position statement by the Canadian Psychological Association and the Canadian Association of School Psychologists on the Canadian press coverage of the province-wide achievement test results*. Retrieved October 15, 2004 from http://www.cpa.ca/documents/joint_position.html

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education*, 51(5), 334-344.

Stone, J.E. (1999). Value added assessment: An accountability revolution. In M . Kanstoroom and C. E. Finn, Jr. (Eds.), *Better teachers, better schools*. 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.

Volante, L (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, 35 (September 25). Retreived December 15, 2004 from http://www.umanitoba.ca/publications/cjeap/articles/volante.html

Wickstrom, R. A. (1999). Accountability: Toward an appropriate approach. *Education Canada*, 39(3), 16-19.

Wolfe, R., Childs, R., & Elgie, S. (2004). *Final report of the external evaluation of EQAO's assessment processes*. Retrieved 11/01/2004 from EQAO Web site: http://www.eqao.com/pdf_e/04/04p014e.pdf

Wolfe, R., Wiley, D., & Traub, R. (1999). *Psychometric perspectives for EQAO: Generalizability theory and applications*. Retrieved 11/01/2004 from EQAO Web site: http://www.eqao.com/pdf_e/99/99P033e.pdf

Wright, S. P. (2004). *Advantages of a multivariate longitudinal approach to educational value-added assessment without imputation*. Retrieved December 15, 2004 from http://www.wmich.edu/evalctr/create/2004/Wright-NEI04.pdf