Improving the Utility of Large-Scale Assessments in Canada

W. Todd Rogers
University of Alberta

Abstract

Principals and teachers do not use large-scale assessment results because the lack of distinct and reliable subtests prevents identifying strengths and weaknesses of students and instruction, the results arrive too late to be used, and principals and teachers need assistance to use the results to improve instruction so as to improve student learning. Therefore, it is recommended that the first assessment activity should be to clearly establish that the domain to be assessed is multidimensional. Given this, the assessment schedule should be changed so that a given subject area is assessed in non-consecutive years but the number of sittings remains the same each year. Assistance should be provided to principals and teachers so as to increase their understanding of how to use large-scale assessment results. Three suggested assessment cycles are presented, each of which increases the reliability of subtests and provides principals and teachers with at least two years to make changes in instruction.

Keywords: large-scale assessments, subtests, issues, solutions, multidimensional

Résumé

Les directeurs et enseignants n'utilisent pas les résultats d'évaluations à grande échelle car le manque de sous-tests fiables et distincts empêche l'identification des forces et faiblesses des étudiants et de l'instruction, les résultats arrivent trop tard pour être utilisés et les directeurs ainsi que les enseignants ont besoin d'aide pour exploiter les résultats afin d'améliorer l'instruction, de manière à améliorer l'apprentissage des élèves. Il est donc recommandé que la première activité d'évaluation consiste à établir clairement que le domaine à évaluer est multidimensionnel. Ceci étant établi, le programme d'évaluation devrait être modifié de manière qu'un domaine donné soit évalué lors d'années non consécutives, mais que le nombre de séances demeure identique chaque année. Une aide doit être apportée aux directeurs et enseignants afin d'accroître leur compréhension de la façon d'utiliser les résultats d'évaluations à grande échelle. Trois cycles d'évaluation suggérés sont présentés, chacun augmentant la fiabilité des sous-tests et donnant aux directeurs et enseignants au moins deux ans pour apporter des modifications aux méthodes d'instruction.

Mots-clés : évaluations à grande échelle, sous-tests, problèmes, solutions, multidimensionnel

This article was first presented as a Presidential Address delivered to members of the Canadian Educational Researchers' Association (CERA) during the annual conference of the Canadian Society for the Study of Education (CSSE), in Victoria, June 3, 2013.

Cet article fut présenté pour la première fois sous la forme d'un message du président aux membres de l'Association canadienne de chercheurs en éducation (ACCE) lors de la conférence annuelle de la Société canadienne pour l'étude de l'éducation (SCÉÉ), à Victoria, C.-B., le 3 juin 2013.

Introduction

Government policy makers and officials in Canada believe public evidence of student performance drawn from sound and credible large-scale assessments will help focus educators' attention on improving curriculum and instruction and, as result, enhance student learning and performance. While there are many who agree that data-informed decision making is a way to improve curriculum and instruction so that student learning is improved, others question the perceived value of large-scale assessments as a way to improve curriculum and instruction so that student learning is improved. The purpose of this article is to briefly review the beneficial and detrimental effects of large-scale testing, discuss three existing concerns, and propose a change in the scheduling of large-scale assessments to properly address the three concerns.

Beneficial Effects of Large-Scale Assesments

Briefly, students across schools are treated equitably and fairly by providing a common "yardstick" in the form of a common assessment (Phelps, 2008; O'Conner, 2009). Large-scale assessment results have positively affected the need for increased attention to students with special needs (Roderick & Engel, 2001; Thurlow & Ysseldyke, 2001). Providing assessment results to students, be they classroom assessments or large-scale assessments, has a strong positive effect on student achievement (Phelps, 2012). After being involved in item writing, item review panels, sensitivity review panels, and scoring students' responses to open-ended items, principals and teachers return to their schools and classrooms with enhanced training and experience in item writing and scoring that they can apply to their own classroom assessments (Cizek, 2001). Further, they can use large-scale assessment results to identify the need for professional development presentations and workshops (Cizek, 2001). Members of departments of education and school district personnel can use large-scale assessment results to confirm that the curriculum has been addressed effectively (Lissitz & Schafer, 2002). The presence of publicly reported large-scale assessment results can serve as a conversation starter for a discussion about what should constitute an accountability system, the implementation of which is vital if the goal is to improve education and achievement of students (Cizek, 2001; Ferrera, 2005; Mirazchiyski, 2013; Paton, 2013).

Detrimental Effects of Large-Scale Assesments

Among the points of concern for large-scale assessments are that large-scale assessments narrow instructional content with a concomitant emphasis on students learning lower order thinking skills at the expense of higher order thinking skills; reduce instructional time in favour of test preparation activities; yield results for students who are no longer the teachers' students since the students move to the next grade or to a junior or senior high school; increase cheating; and reduce teacher professionalism (Brandt, 1995; Burrows, Groce & Webeck, 2005; Chester, 2005a, 2005b; Darling-Hammond, Ancess, & Falk, 1994; Earl & Katz 2006; Kohn, 2000; National Council on Measurement in Education, 2012; Popham, 2002; Shepard, 1991, 2010; Wiggins & McTighe, 2005). These concerns continue to be expressed and, with the introduction of programs like No Child Left Behind (United States Department of Education, 2003) and Race for the Top (United States Department of Education, 2009), they now include restrictions on teacher authority; questionable evaluations of school personnel and teacher stress; unwarranted reductions of teacher salaries; school sanctions; neglect of content not covered by the assessments (e.g., if science is not assessed, then perhaps science is not all that important to learn); inconsistent performance standards and cut-scores across grades and over years; and inconsistent school results over time (e.g., Berliner, Popham, & Shepard, 2000; Burrows et al., 2005; Chester, 2005a; Childs & Fung, 2009; Cizek, 2001; Kane & Staiger, 2002; Klinger & Rogers, 2011; Klinger, Shula, & Wade-Wooley, 2009; Linn, 2003; Thompson, 2001). Ravitch (2010) compellingly summarizes these concerns, concluding that curriculum and instruction are far more important than large-scale assessments and that testing has become an end in itself and not a means to the end.

The Situation in Canada

The majority, but not all, of the research cited above was conducted in the United States. While a similar situation exists in Canada, there are two important differences between the provinces/territories in Canada and all but a few states (e.g., Hawaii, Texas) in the United States. First, the curriculum for a subject area in each province/territory is common to all schools in the province. Teachers in all schools in the province/territory must provide learning opportunities to their students to enable them to learn the knowledge

and acquire the thinking, problem solving, and reasoning skills identified in the learning expectations provided in the program of studies or curriculum guide. Consequently, large-scale provincial/territorial assessments can be used to improve instruction but not the curriculum at the school level. Second, the sanctions that exist in the United States are not present to the same degree in Canada. While some teachers may elect not to teach a grade that has a provincial/territorial assessment, accountability in Canada is commonly framed within the context of professional responsibility, with the expectation that principals and teachers will use the results of large-scale assessments to inform and support their own ongoing school improvement efforts to improve student learning and performance.

However, the contention in this article is that large-scale assessment results in Canada and elsewhere will become more useful to the extent that they provide the following:

a. relevant and justifiable evidence to foster a conversation about how to improve instruction of all students;
b. adequate time for principals and teachers so that they can meaningfully engage in sound and valid planning and implementation of needed instructional changes; and
c. assistance to teachers and principals to help them use the information from large-scale assessment and to integrate the information with information gained from their own classroom assessments.

It is argued in this article that these three points can be effectively addressed by providing credible diagnostic information and more time for and assistance to principals and teachers to allow them to use these this information to improve instruction in ways that enhance student learning and achievement.

Three Issues That Need To Be Addressed

Lack of Credible Diagnostic Information

To start a conversation about how to improve instruction for all students, principals and teachers need reliable "diagnostic" information that they can validly interpret in terms of strengths and weaknesses of their students. They clearly recognize that subtest scores will

allow them to see what changes are needed to improve what is taught and how it is taught so as to improve their students' learning (Hattie & Timperley, 2007). As indicated above, providing feedback to students that they can use to identify their own strengths and weaknesses leads to improved student learning and performance (Phelps, 2012).

The premise for reporting subtest scores from curriculum-based assessments is based on the assumption that the curriculum is multidimensional. For example, most mathematics curricula are divided into five or so content subdomains such as number sense and numeration, measurement, geometry and spatial sense, patterning and algebra, and data management and probability. As well, most curricula include a cognitive component such as knowledge and application, problem solving, reasoning, and evaluation. Clearly, the names convey differences among the content subdomains and the cognitive levels. But are the content and cognitive subdomains *distinct* or *not*?

To answer this question, the *first* activity in the test development process should be to clearly establish *domain clarity*. Deliberate effort needs to be devoted at the very beginning of the assessment process to determine if the domain to be assessed is unidimensional or multidimensional. If the domain is found to be unidimensional, then there is no warrant to report subtest scores. If the domain is found to be multidimensional, then there is a warrant to report subtest scores. However, what most frequently happens is that an implicit assumption is made that the curriculum is multidimensional (Haladyna & Kramer, 2004). Items are carefully developed for each subdomain such that each item is relevant to the subdomain and the set of relevant items represents the subdomain (Messick, 1989). But the total number of items for each subdomain is limited so that the full test can be administered in two to three hours, given the age of the students to be assessed.

Despite the small number of subtest items, the question "Can we report subscores?" often arises after the assessment has been administered. At this point, methods for empirically determining if subtest scores are distinct or if they add value over the total

score are used.[1] Generally, the findings reveal that reporting subtest scores for current large-scale assessments is not warranted. The common reasons are high subtest correlations and low subtest reliabilities. Greg Cizek, in his National Council of Measurement and Evaluation Presidential Address (April 29, 2013), firmly stated that reporting subtest scores from current assessment instruments and tests was simply *inappropriate*.

What is wanted by principals and teachers is displayed in Figure 1. The five subdomains have been determined to have a good amount of uniqueness by a panel of subject matter experts who know well the knowledge and skills of the domain and its subdomains to be learned and the characteristics of the students to be taught. Further, the subtests developed to measure each of the subdomains are composed of relevant and representative items. First consider Students A and B. The performance of Student A across the five parts of the curriculum is uniformly low and the performance of Student B across the five parts of the curriculum is uniformly high. Students like Students A and B learn the knowledge and skills of each subdomain equally well, but at different levels of performance. Students like Student C do not learn the knowledge and skills of each subdomain equally well. Teachers can use the profiles of students like Student A to develop a full remedial plan to help erase their general low performance and a remedial plan for students like Student C to address the subdomains with low performance while maintaining their performance for the subdomains with high performance. The point to make here is that the profiles displayed in Figure 1 can only be obtained if it is clearly and well established that the domain is multidimensional to at least some degree before items are developed. If the domain is not multidimensional, then profiles like that shown for Student C will not be realized.

1 Among the methods used to determine if subtest scores are distinct are (a) correlations corrected for attenuation (Haladyna & Kramer, 2004; McPeek, Altman, Wallmark, & Wingersky, 1976), (b) agreement method (Babenko & Rogers, 2014; Kelly, 1923; Gulliksen, 1950; Lord & Novick, 1968), and (c) generalizabilty analysis (Rogers & Radwan, 2012). The proportional reduction in mean square error (Haberman, 2005; Sinharay, Haberman, & Puhan, 2007; Sinharay, Puhan, & Haberman, 2009; Sinharay, 2010) can be used to determine if a subtest score adds value over the total test scores.

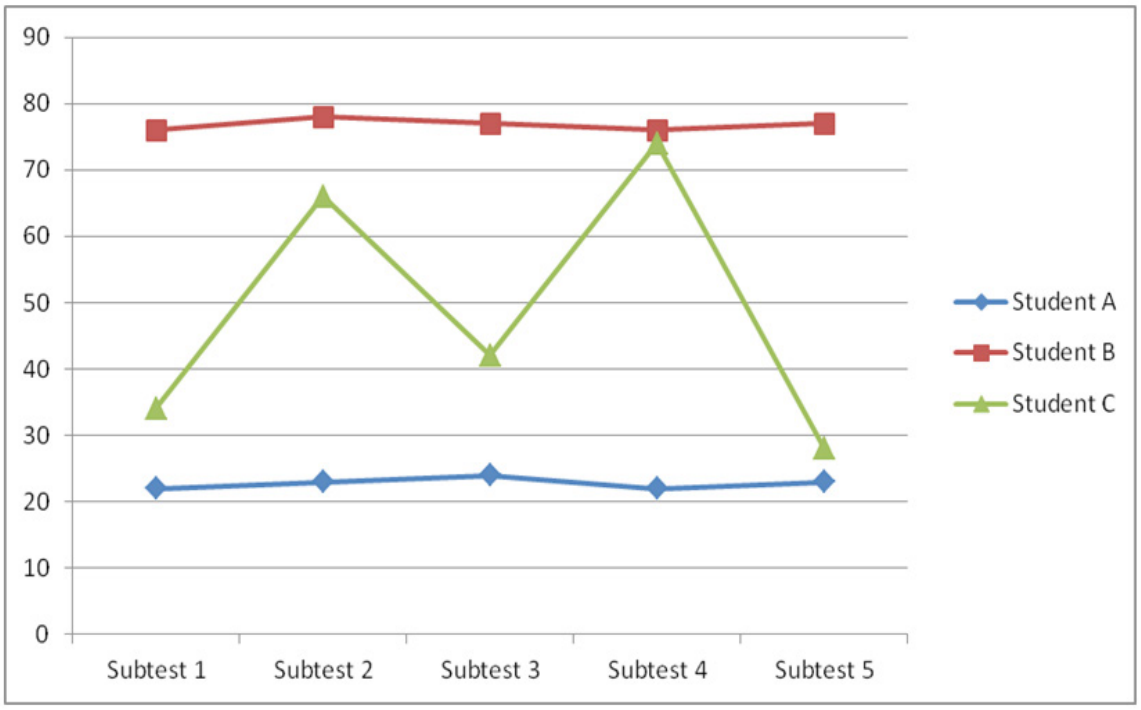

Figure 1: Sample Student Profiles for Five Parts of the Curriculum

Now, what has been said in the previous paragraph is predicated on the assumption that the subtests are reliable. But the reliabilities of subtests developed to measure subdomain performance are generally too low (Haladyna & Kramer, 2004; Luecht, Brumfield, & Breithaupt, 2006). Time limits for the administration of the total assessment instrument are set according to the age of the students and rarely exceed two hours, with perhaps an extra half hour as an accommodation for students who need more time. Consequently, the number of items for the subtests included in an assessment instrument is small, which in turn limits the reliability of the subtests. A sufficient number of items need to be included in each subtest so that the reliability of each subtest is adequate to report subtest scores. In all of the cases in which investigations have been conducted to see if subscore reporting is warranted, the main emphasis has been on developing a set of items for the full assessment that represents the full domain according to the table of specifications for the full assessment and with a two-hour test administration limit.

Therefore, to warrant reporting of profiles of subtest scores, consideration needs to be given to three conditions when constructing assessment instruments. The first, a substantive condition, is that the construct or domain to be assessed is multidimensional.

The second and third conditions are statistical in nature, namely, that there are low correlations among subtests and high subtest reliabilities.

Lack of Adequate Time to Profitably Use the Results

Large-scale assessments are generally administered toward the end of the school year or, for semestered schools, the end of the semester. The staff of test agencies responsible for scoring students' responses to open response items, analyzing the scored responses, equating one assessment to another within and/or across years, and preparing reports work diligently during July and the first half of August to get results to schools before the beginning of the next school year.

Despite this effort, *can principals and teachers do what they are supposed to do before school starts*? They need to do each of the following before classes begin:

- interpret the results from the large-scale assessments;
- integrate what they have learned with what they know from their own classroom assessments to identify strengths and weaknesses in student learning;
- review what they did during the last year to allow the students to acquire the knowledge and skills their students were expected to learn; and
- identify and make sound and credible changes to their teaching materials, activities, and instructional approach that will enhance the learning and achievement of the students they have for the coming year.

Typically, teachers are expected back to school during the week before school starts or on the first day of school. Given the startup activities principals and teachers are responsible for at the beginning of the school year, they have little or no available time to use large-scale assessment results in a meaningful way and do what they need to do to improve instruction before the school year begins. Consequently, they often simply ignore the large-scale assessment results (Klinger & Rogers, 2011).

Lack of Needed Assistance

Principals and teachers need assistance with using large-scale assessment results in a meaningful way. They need assistance with interpreting the assessment results for their school, integrating the results with what they know from their own classroom

assessments, and how to use the combined results to identify strengths and weaknesses in the teaching materials, activities, and instructional approaches they use at the class level (Deluca & Klinger, 2010; Webber, Aitken, Lupart, & Scott, 2009). Teachers' lack of understanding of the purposes and uses of large-scale assessments has a negative impact on their use of and attitude toward large-scale assessments (Aiken, 1991; Bracey, 2005; Burger & Krueger, 2003; Cannell, 1987; Earl, 2003; Fairhurst, 1993; Lewis, 2007; Smith, 1991).

Proposed Solutions

Ensure Relevant and Trustworthy Diagnostic Evidence

In order to find students with profiles like Student C in Figure 1, it is necessary that the curriculum be truly multidimensional. This requires that the parts of the curriculum are distinct and not highly correlated and the subtests developed to assess each part are composed of relevant and representative items, have adequate reliability, and do not correlate highly. Thus, the first step to take when developing an assessment is to carefully and deliberately consider whether or not the different subdomains of the curriculum are distinct. It may well be that the subdomains are related, but there must be some uniqueness for each subdomain to warrant the claim of distinction. Members of a panel of experts in the subject area who are knowledgeable about the students to be assessed should independently determine what makes each subdomain unique and then reach consensus. The uniqueness might include content and/or thinking skills that are needed to learn the content and acquire the skills for each subdomain.

If the subdomains of the curriculum are judged to be at least partially distinct and items relevant to and representative of each subdomain have been constructed, then responses from a representative sample of students gathered from a pilot study or field trial can be analyzed to provide empirical evidence that the domain is multidimensional. The responses of the students to each item in a subtest *should* correlate to the following:

a. highly with the subtest score (i.e., high item discrimination within subtest); and
b. lowly with the subtest scores from the other subtests (i.e., low item discrimination across subtests other than the subtest the item belongs to).

Given a. and b. are met, then

c. the correlations among the subtest scores should be lower than the reliabilities of the subtests to be correlated.

Change the Assessment Schedule

The present schedule of annually assessing students in the same subject areas should be changed to allow sufficient time for principals and teachers to use subtest scores to plan what changes in instruction may be needed and to implement the changes in a reasonable and steady manner. To achieve this, the assessment schedule should be changed so that a given subject area is not assessed every year but the total number of times a student sits for the assessments remains the same or essentially the same—three to four sittings of 2 hours each. Adoption of this proposal will

1. allow principals and teachers more time to integrate the large scale assessment results with their own classroom assessment results, identify areas of strength and weakness, review what they did the previous year, and formulate and implement changes to their teaching materials, activities, and instructional approach to address weaknesses while maintaining strengths,

and, at the same time

2. allow a greater number of items for each subtest in order to ensure high subtest reliability.

The three possible options that follow are provided to illustrate possible administration schedules.

One of three assessments per year. If three subject areas (e.g., either mathematics, reading, and writing, or literacy, mathematics, and science) are assessed each year, then the change would lead to one subject area being assessed each year in three sittings. Three years would be needed to complete a cycle as shown in Figure 2. Three sittings per year would triple the number of subtest items, thereby leading to an increase in subtest reliability. Principals and teachers would have a greater period of time between two consecutive assessments of the same subject area to integrate the large-scale assessment results with their own classroom assessment results, identify strengths and weaknesses, and formulate changes to be made to the teaching materials, activities, and instructional approach,

perhaps during the first year. During the second year, they could try out the changes and make revisions as needed, and then implement the revisions in the third year.

In the fourth year, the first subject area would be assessed again. Using the results from this second assessment, principals and teachers could be properly held accountable for the changes to the teaching materials, teaching activities, and/or instruction they had made in an attempt to improve student performance in their school between Year 1 and Year 4.

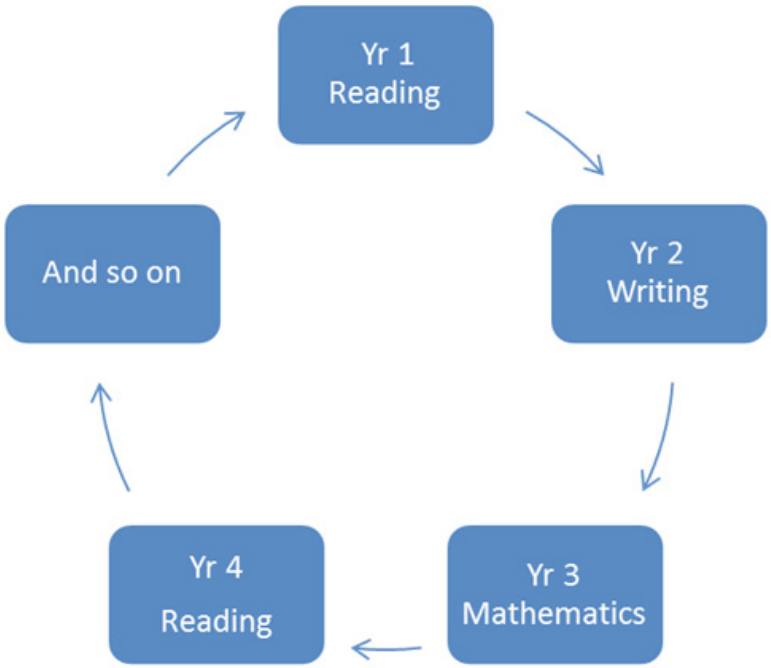

Figure 2: Assessment of One Subject Area Each Year

Two of three assesments on year and one assesment the next year. Some may argue that three years between between the assessment of a subject area, two subject areas would be assessed in two or three sittings for each assessment in the first year and one subject area would be assessed in two or three sittings in the second year (see Figure 3). The advantage of the second option is that each subject area would be assessed every two years instead of every four years as in the first option (cf. Figures 2 and 3). The planning and implementation stage would be shortened to two years—planning and pilot testing during

the first year and full implementation in the second year.

A potential disadvantage of this option is that the students would take two different assessments one year and one assessment in the next year. A further disadvantage is the number of items for each assessment would be reduced by a third if two sittings were used for each assessment, which could result in subtest reliabilities that are too low.

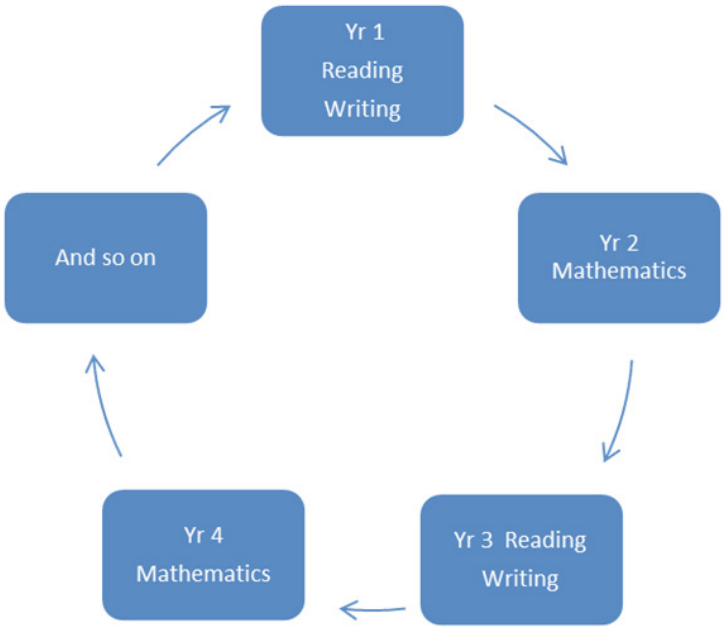

Figure 3: Assessment of Two Subjects in One Year and a Third Subject in the Next Year

Two assessments each year. In some jurisdictions as many as four different subject areas are assessed in one year. The schedule in this case would look like what is presented in Figure 4 for literacy (reading and writing combined), mathematics, science, and social studies. Each subject area would be assessed every two years. The number of sittings for each subject area would be at least two. The advantage of this option is that each subject area would be assessed every two years as in the previous option (Figure 3) instead of every five years (Figure 2 with four assessments). As with the second option, a potential disadvantage of the third option is the number of items for each assessment would be reduced by a third if two sitting were used for each assessment, which could result in

subtest reliabilities that are too low.

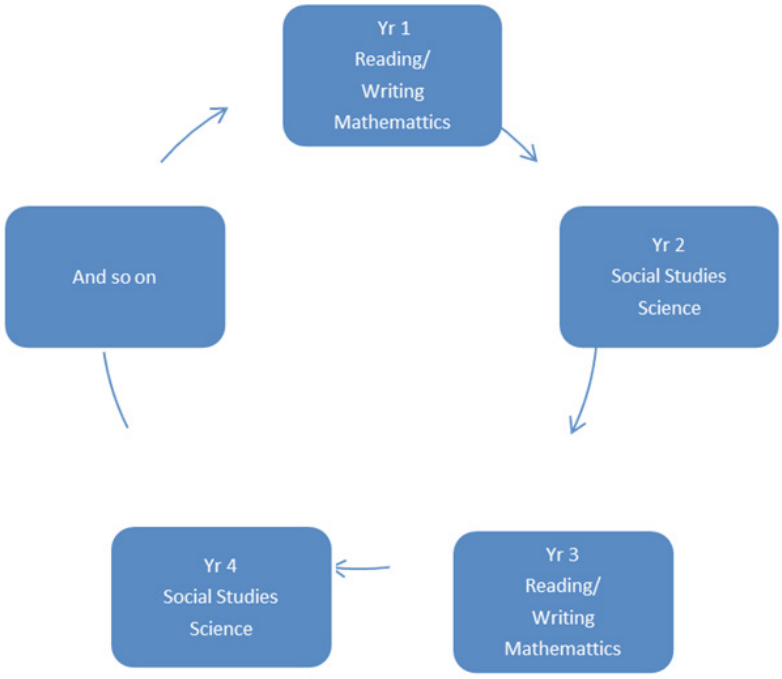

Figure 4: Assessment of Two Subjects Each Year

A potential disadvantage of all three schedules is that teachers would concentrate more on the subject area(s) to be assessed that year and less on the non-assessed subject areas. This disadvantage could be addressed by ensuring that teachers follow through with planning and implementation of the changes identified from the previous year's assessment(s).

Increased Reliability

At different points an increase in reliability has been mentioned. If the first option (Figure 2) were to be adopted, then the number of items would triple. If either the second (Figure 3) or third (Figure 4) options were adopted, then the number of items would at least double. Based on the subtest reliabilities observed for different assessments conducted today, subtest reliabilities typically range from about 0.50 to 0.80, with a median and mean

close to 0.65. As shown in Table 1, application of the Spearman-Brown formula (Lord & Novick, 1968, p. 112) would yield a range from

1. 0.66 to 0.89 with a median and mean close to 0.79 if the number of items for each subtest were doubled

and

2. 0.75 and 0.92 with a median and mean of 0.85 if the number of items in each subset were tripled.

Table 1: Subtest Reliabilities if Double or Triple Number of Items

Initial	Test Length	
Reliability	Double	Triple
0.50	*0.67*	*0.75*
0.60	*0.75*	0.82
0.65	0.79	0.85
0.70	0.82	0.88
0.80	0.89	0.92

Reliabilities below 0.80 are perhaps too low, but if the recommendation given earlier to select items for a subtest for which the student responses correlate highly with the subtest score and lowly with the other subtest scores is accepted, then adequate reliability would likely be obtained.

Provide Adequate Support

Principals and teachers need assistance with interpreting and using large-scale assessment results for their schools. Personal assistance provided by members of an assessment agency's outreach team throughout the year and interactive online reporting mechanisms, where principals and teachers can compare the performance of their schools with schools with similar bio-demographic characteristics, are two ways the needed assistance can be continuously provided. Although teachers can use documents like the *Student Evaluation*

Standards (Joint Committee on Standards for Education Evaluation, 2003) and the *Principles for Fair Student Assessment Practices for Education in Canada* (Centre for Research in Applied Measurement and Evaluation, 1993) to assist them to improve their own assessment practices, many do not, likely because of a lack of relevant knowledge and skills and confidence. There clearly is a need to support principals and teachers to use large-scale assessment results in a reasoned and effective way to improve student learning and performance (Webber et al., 2009).

Conclusion

Large-scale assessments have assumed the preeminent role in educational accountability and reform because they provide a common and, ostensibly, a fairer yardstick to monitor student achievement over time and to compare schools and school districts. They are relatively efficient and, importantly, the results are visible (Linn, 2000). But, with one or two exceptions (e.g., Japan), have we seen much change in student performance and are the results used as expected, particularly at the local school level? To correct this situation, we strongly recommended the following be considered:

1. it be clearly confirmed that the curriculum is indeed multidimensional;

and, if confirmed, that

2. procedures for item analysis be expanded to include both the subtest to which an item is referenced and the subtests to which the item is not referenced;
3. the assessment schedule be changed to
    a. provide more time for educators to review the results to identify areas of strength and areas of weakness, to formulate changes to address weakness while maintaining strength, acquire any needed materials, and implement the changes in a reasonable and thoughtful way, and
    b. allow a greater number of items for each subtest, thereby increasing subtest reliabilities;

and

4. assistance be provided to school principals and teachers to help them work with the large-scale assessment results and the knowledge they have about their own instruction during the last year to make changes so as to increase student learning

and achievement.

Principals and teachers must be provided with reliable profiles that can be validly interpreted, and they must have adequate time and assistance to make needed changes to enhance learning and achievement of all of their students.

References

Aiken, L. R. (1991). Detecting, understanding, and controlling for cheating on tests. *Research in Higher Education*, New York, NY: Human Sciences Press, *32*(6), 725–736.

Babenko, O., & Rogers, W. T. (2014). Comparison and properties of correlational and agreement methods for determining whether or not to reports scores. *International Journal of Learning, Teaching and Educational Research.*

Berliner, D. C., Popham, W. J., & Shepard, L. A. (2000, April). *Three blueprints for a revolution: How to halt the harm caused by high-stakes tests.* General Session presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bracey, G. W. (2005). The 15th Bracey report. *Phi Delta Kappan, 87*(2), 138–153.

Brandt, R. (1995). Punished by rewards: A conversation with Alfie Kohn. *Educational Leadership, 53*(1). Retrieved from Alfie Kohn website: http://www.alfiekohn.org/teaching/pbracwak.htm

Burger, J. M., & Krueger, M. (2003). A balanced approach to high-stakes achievement testing: An analysis of the literature with policy implications. *International Electronic Journal for Leadership in Learning, 7*(4). Retrieved from University of Calgary website: http://www.ucalgary.ca/~iejll/

Burrows, S., Groce, E., & Webeck, M. L. (2005). Social studies education in the age of testing and accountability. *Educational Measurement: Issues and Practice, 3,* 13–20.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average.* Daniels, WV: Friends for Education.

Centre for Research in Applied Measurement and Evaluation. (1993). *Principles for fair student assessment practices for education in Canada.* Retrieved from University of Alberta website: http://www.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf

Chester, M. D. (2005a). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice, 4,* 40–52.

Chester, M. D. (2005b). Measuring the impact of state accountability programs. *Educational Measurement: Issues and Practice, 4,* 3–4.

Childs, R. A., & Fung, L. (2009). "The first year they cried": How teachers address test stress. *Canadian Journal of Educational Administration and Policy, 64.* Retrieved from University of Manitoba website: http://www.umanitoba.ca/publications/cjeap

Cizek, G. J. (2001).More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Darling-Hammond, L., Ancess, J., & Falk, B. (1994). *Authentic assessment in action: Studies of schools and students at work.* New York, NY: Teachers College Press.

Deluca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidate's learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419–438.

Earl, L. M. (2003). *Assessment as learning.* Thousand Oaks, CA: Corwin Press.

Earl, L., & Katz, S. (2006). *Leading in a data rich world: Harnessing data for school improvement.* Thousand Oaks, CA: Corwin Press.

Ferrara, D. (Ed.). (2005). [Special issue]. *Educational Measurement: Issues and Practice, 24*(4).

Fairhurst, D. (1993). Achievement tests and dissonance. I*ssues, Events, and Ideas, Early Childhood Council of the Alberta Teachers' Association, 70,* 8.

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: John Wiley.

Haberman, S. J. (2005). When can subscores have value? (ETS Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service.

Haladyna, T. M. & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 27,* 349–368.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Joint Committee on Standards for Education Evaluation. (2003). *The student evaluation standards: How to improve the evaluation of students.* Thousand Oaks, CA: Corwin Press.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 235–283). Washington, DC: Brookings Institute.

Kelley, T. L. (1923). A new method for determining the significance of differences in intelligence and achievement scores. *Journal of Educational Psychology, 14,* 300–303.

Klinger, D. A., & Rogers, W. T. (2011). Teachers' perceptions of large-scale assessment programs within low-stakes accountability frameworks. *International Journal of Testing, 11,* 1–22.

Klinger, D. A., Shula, L. A., & Wade-Wooley, L. (2009). Towards an understanding of gender difference in literacy achievement. Toronto, ON: Education Quality and Accountability Office. Retrieved from EQAO website: http://www.eqao.com/Research/pdf/E/FINAL_ENGLISH_Gender_Gap_Report_As_of_May_11_2010.pdf

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools.* Portsmouth, NH: Heinemann.

Lewis, A. C. (2007). How well has NCLB worked? How do we get the revision we want? *Phi Delta Kappan, 88*(5), 352–353.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 23*(9), 4–14.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(3), 3–13.

Lissitz, R. W., & Schafer, W. D. (2002). *Assessment in educational reform.* Boston, MA: Allyn and Bacon.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* New York, NY: Addison–Wesley.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for the Uniform CPA Examination. *Applied Measurement in Education, 19,* 189–202.

McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test (GRE Board Professional Report No. 74–4P). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED163090)

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

Mirazchiyski, P. (2013). *Providing School- Level Reports from International Large-Scale Assessments: Methodological Considerations, Limitations, and Possible Solutions.* Amsterdam: International Association for the Evaluation of Educational Achievement.

National Council on Measurement in Education. (2012). *Testing data integrity in the administration of statewide student assessment programs.* Madison, WI: Author.

O'Connor, K. (Ed.). (2009). *How to grade for learning, K-12*. Thousand Oaks, CA:Corwin Press.

Paton, P. (2013, March 31). Girls "marked up in lessons to reward good behavior." *The Telegraph (London).* Retrieved from The Telegraph website: http://www.telegraph.co.uk/education/educationnews/9963834/Girls-marked-up-in-lessons-to-reward-good-behaviour.html

Phelps, R. (2008). The role and importance of standardized testing in the world of teaching and training. *Nonpartisan Education Review, 4*(3), 1–9.

Phelps, R. (2012). The effect of testing on achievement: Meta-analysis and research summary, 1910–2010. *International Journal of Testing, 12,* 21–43.

Popham, W. J. (2002, April). *High-stakes tests: Harmful, permanent, fixable.* Paper presented at the Annual Conference of the American Research Council, New Orleans, LA.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education.* New York, NY: Basic Books.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low achieving students to high-stakes testing. *Educational Analysis and Policy Analysis, 23*(3), 197–227.

Rogers, W. T., & Radwan, N. (2012). Use of Generalizability Theory to determine if subtests are distinct. Report submitted to the Education Quality and Accountability Office. Toronto, ON.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(6), 2–16.

Shepard, L. A. (2010).What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education, 85,* 246–257.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150–174.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26,* 21–28.

Sinharay, S., Puhan, G., & Haberman, S. (2009, April). Reporting diagnostic scores: Temptations, pitfalls, and some solutions. Paper presented at the National Council on Measurement in Education, San Diego, CA.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8–11.

Thompson, S. (2001). The authentic testing movement and its evil twin. *Phi Delta Kappan, 82*(5), 358–362.

Thurlow, M. L., & Ysseldyke, J. E. (2001). Standard setting challenges for special populations. In G. J. Cizek (Ed.), *Setting performance standards: Concepts,*

methods, and perspectives (pp. 387–410). Mahwah, NJ: Lawrence Erlbaum Associates.

United States Department of Education. (2009). *Race to the Top program: Executive summary.* Washington, DC: US Department of Education. Retrieved from United States Government website: https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

United States Department of Education. (2003). No child left behind, accountability and adequate yearly progress (AYP). Washington, DC: US Department of Education. Retrieved from United States Government website: http://www2.ed.gov/nclb/landing.jhtml

Webber, C. F., Aitken, N., Lupart, J., & Scott, S. (2009). *The Alberta student assessment study: Final report.* Edmonton, AB: Alberta Education.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design.* (Expanded 2nd ed.). Upper Saddle River, NJ: Merrill/ASCD.