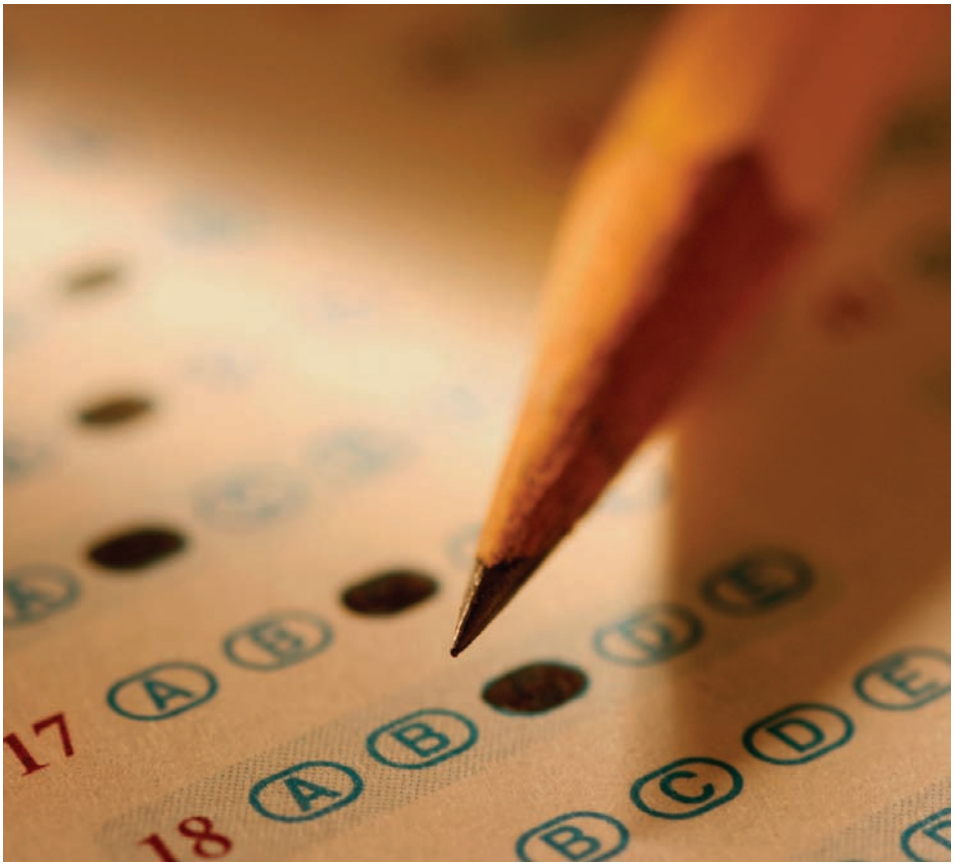
Testing achievement

Louis Volante says standardised achievement testing and school improvement is a zeitgeist in need of correction

Introduction

Governments and senior educational policy-makers in much of the western world have explicitly endorsed the philosophy that test-based accountability represents the best way to spur improvements in schools. By holding teachers accountable for external test performance, educators are compelled to improve their classroom practice, thereby raising the educational performance of all students and also narrowing the gap between low and high achieving students.

The widespread endorsement of this philosophical orientation is akin to an educational zeitgeist – the characteristic thought, preoccupation or spirit of a particular period in time. This article examines some of the central rationales driving test-based accountability models (also referred to as standards-based reform) and discusses the necessity of shifting to data-integrated decision-making that effectively balances different forms of student assessment for accountability purposes.



Assumptions guiding standards-based reform

There are a variety of interrelated rationales that drive standards-based reform movements around the world. Rather than discuss the merits and limitations of each of these assumptions, I focus on three key propositions that underlie the increased use of standardised testing to drive school improvement:

1 High-stakes testing is the most effective policy-lever to improve teaching and learning;
2 A focus on getting back-to-basics by testing key curriculum areas such as language arts, mathematics, and potentially science, is the most effective way to gauge educational quality and to improve schools; and
3 Standardised test results are the most reliable and valid form of data for planning and improvement purposes. Closer scrutiny of each of these rationales suggests the need to re-examine the basis of contemporary standards-based reform initiatives.

Policy-lever rationale

The chief rationale driving standards-based reform is that high-stakes testing is the most effective lever to improve teaching and learning within schools. Regional and/or national testing targets provide benchmarks to gauge educational progress and provide the necessary stimulus to spur school improvement. This assertion, however, is not supported by the empirical literature. In fact, four large reviews on the impact of formative assessment have demonstrated that these practices double the speed of student learning (Black & Wiliam, 1998; Crooks, 1988; Kluger & DeNisi, 1996; Natriello, 1987).

Improved formative assessment helps struggling students more than others and thus reduces the achievement gap while still raising achievement overall. Thus, formative assessment, not summative standardised testing, provides teachers with the most salient data to inform their practice and improve student learning and achievement.

Although it is difficult to draw comparisons across nations with diverse student populations, current trends suggest that the intensification of external testing for accountability purposes has not led to significant improvements in achievement. For example, countries with the most intensified external testing programs such as England and the United States have not fared well on international measures such as the *Trends in International Mathematics and Science Study* (TIMSS), *Progress in International Reading Literacy Study* (PIRLS), and *Programme for International Student Assessment* (PISA).

Ironically, countries such as Finland, which possess no centrally administered testing measures, seem to be leading the pack in the previous international assessments. These findings dispel the notion that countries which emphasise external testing are more likely to raise student achievement.

Back-to-basics rationale

Another important rationale used to support standards-based reform is that a focus on key curriculum areas such as language arts and mathematics is essential for school improvement. Other non-core subject areas such as music, physical education, visual arts, etc. are too costly and difficult to assess and may actually provide a distraction from what should be a focus on getting back-to-basics. The end result is that schools inevitably become oriented toward dual-focused (language arts, mathematics) or tri-focused goals (language

arts, mathematics, science) versus multi-focused in their attention to the mandated curriculum.

Research across various countries suggests that this becomes especially problematic when high-stakes consequences are attached to test performance (Black & Wiliam, 2005; Haertel & Herman, 2005). The inevitable reaction is 'teaching to the test' with non-tested subject matter squeezed out of the curriculum. When one considers the long-standing criticism that most standardised measures focus on low-level thinking skills, the prospect of developing a critical mass of students ready to spur innovations in the knowledge economy is greatly diminished.

Reliability and validity rationale

The advent and widespread use of standardised testing measures is partly a response to the perception that teachers' classroom assessment data are more subjective and less reliable and valid than standardised measures. In truth, part of this assertion is currently unassailable – classroom-based evaluative judgments can vary significantly between teachers. Nevertheless, this seems unlikely to change in systems that put resources and excessive energy in external testing programs.

The underlying message for teachers working within such systems is that their judgments are inconsequential for significant decision-making purposes and they don't necessarily have to improve their current assessment literacy – an understanding of the principles and practices of sound assessment (Stiggins, 2004). If classroom assessment is to have more robust reliability, then practicing teachers require ongoing professional development opportunities to develop their assessment expertise in a context that values those judgments (Volante, 2006).

The second part of this assumption deals with the notion of validity, which is generally defined as the accuracy of assessment-based interpretations. It is difficult to concede that external testing is the most valid mechanism for gauging authentic student learning when one considers the limitations of paper-and-pencil standardised testing measures and the richness of many classroom performance-based tasks (i.e. presentations, projects, experiments). Rather, what standardised measures gain in reliability they often lose in test validity.

A student is not necessarily literate if they score extremely high on a standardised reading and writing measure, but lack speaking and listening skills that can only be assessed through more authentic tasks. Thus, our concern for the reliability of teachers' evaluative judgments must be tempered by the realisation that particular classroom assessment tasks are vital to provide a comprehensive picture of student and school success.

Moving forward with data-integrated decision-making

Data-driven decision-making has become an increasingly important concept with the multitude of assessment information that is currently available to administrators and teachers in contemporary schools.

Administrators' and teachers' utilisation of assessment data for planning purposes can take various forms. The first, and lowest level response, typically involves the examination of large-scale results in isolation of other forms of student data (Volante, 2007). Here, teachers and administrators make adjustments to teaching and planning on the basis of general test scores in particular subject areas. This approach often results in large-scale assessment driving, rather than informing, school improvement.

The second level is similar to the first, with the exception that large-scale assessment data is disaggregated for particular student groups (special needs students, English-as-a-second-language students, distinct ability groups, etc). Although the second level shows greater sophistication, it still suffers from the same methodological problem – large-scale assessment drives teaching/planning decisions without the benefit of other forms of student information.

The third, and highest level, involves the integration of disaggregated large-scale assessment results with other forms of student assessment information. Educators at the third level make teaching/planning decisions based on multiple, and at times, contradictory forms of student assessment information. These educators are able to examine student assessment information across a range of subject areas, some which cannot be assessed in a standardised fashion.

Conclusion

In his review of the relationship between standardised testing and school improvement, Raudenbush (2005) concluded that measures of mean proficiency and other statistical analyses of test data alone do not reveal direct evidence of the quality of school practice. This assertion is even more compelling given that it was published by the largest testing agency in North America – Educational Testing Services (ETS).

Clearly, the zeitgeist currently dominating educational reform in much of the western world needs to be corrected. Policy-makers can facilitate this correction by enhancing the salience of classroom-based assessment data for accountability purposes. Interestingly, research has shown how classroom assessment data has been successfully integrated for standards-based accountability purposes in pockets of the United States, England, and Australia (Wilson, 2004). Thus, classroom assessment should be an integral part of an overall accountability framework so that this information is integrated, not isolated, from other forms of student assessment information.

Jurisdictions must develop data-integrated decision-making systems that utilise multiple forms of student assessment data to inform classroom, school, and district planning. In doing so, they provide the public with a more comprehensive approach to assessing school improvement and are well positioned to respond to the complex demands of the 21st century.

This research was funded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

References

Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan, 80(2), 139–48.

Black, P., & Wiliam, D. (2005). Lessons from around the world: How policies, politics and cultures constrain and afford assessment practices. *The Curriculum Journal* 16 (2), 249–261.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58 (4), 438–481.

Haertel, E., & Herman, J. (2005). A Historical Perspective on Validity: Arguments for Accountability Testing (CSE Report 654). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119 (2), 254-284.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist,* 22(2), 155–175.

Raudenbush, S. W. (2004). Schooling, Statistics, And Poverty: Can We Measure School Improvement? Princeton, NJ: Educational Testing Service.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, 86(1), 22–27.

Volante, L. (2006). Reducing bias in classroom assessment and evaluation. Orbit, 36(2), 34–36.

Wilson, M. (Ed.). (2004). Towards coherence between classroom assessment and accountability: 103rd yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press.

Louis Volante Ph.D
Associate Professor,
Brock University, Faculty of Education, Hamilton, Ontario, Canada.
Correspondence to
louis.volante@brocku.ca