

Evaluating Test-Based Accountability Systems:
An International Perspective

by

Louis Volante (Ph.D.)
Assistant Professor
Brock University
Faculty of Education
1842 King Street East
Hamilton, Ontario, Canada L8K 1V7
Email: Louis.Volante@Brocku.ca
Phone: (905) 547-3555 ext. 3621
Fax: (905) 547-9500

Paper Presented at the *Association for Educational Assessment – Europe*
Stockholm, Sweden

November, 2007

Abstract

Calls for greater accountability within schools have led to a rapid expansion of test-based accountability systems within many English-speaking Western nations. Countries such as England, United States, and Canada have all developed educational standards that are primarily measured in relation to standardized achievement test results. Within this Western culture of student performance, this paper discusses the documented impact and social consequences of external testing measures on students, teachers, and school systems. The discussion considers alternative assessment approaches and underscores the importance of shifting to learning-focused accountability.

Standardized achievement tests have been used to measure students' educational progress for nearly a century, but the prevalence of these tests and the accountability purpose they are being asked to serve, have grown substantially during the past two decades (Hamilton, Stecher, & Klein, 2002). The growth of test-based systems within English-speaking Western nations can be partly attributed to their appealing logic: the interplay among content standards, external tests, and accountability is a powerful tool to improve the quality of schools (Abrams, 2004). Widely disseminated student achievement data invites educational leaders, parents, and the broader public to compare and contrast the relative effectiveness of individual teachers, schools, and larger districts. In fixing high stakes to student achievement results, educational policy-makers have borrowed principles from the business sector and attached incentives to learning and sanctions to poor performance on external tests so that schools are forced to become more competitive (Pringle & Martin, 2005). Proponents of this approach routinely argue that test-based accountability will compel teachers to improve their classroom practice, thereby raising the educational performance of all students and also narrow the gap between low and high achieving students. Unfortunately, the fierce debate on the purpose and value of standardized achievement testing, particularly those measures used for high stakes decisions, has often been framed by appealing to rational arguments that are not necessarily grounded in research.

The intent of this paper is to examine and summarize the documented impact and social consequences of test-based accountability systems across various international settings. The concern for the social consequences of testing measures is often referred to as consequential validity (Kane, 2002; Mehrens, 1997; Reckase, 1998; Taleporos, 1998).

Although some individuals in the field of assessment and evaluation disagree with the practice of linking test consequences with the term validity (see Popham, 1997), few would dispute the need to carefully study the impact of external testing measures. If test-based accountability models are to continue in their present structure, the positive consequences should logically outweigh the negative. Mere speculation or conjecture on the part of policy-makers is insufficient; particularly since standardized achievement testing is costly for taxpayers and may be counter-productive to improving national education systems. Essentially, what does the available research suggest are the benefits and drawbacks of utilizing test-based accountability?

Accountability in English-Speaking Western Nations: A Brief Synopsis

The development and implementation of accountability systems has been one of the most powerful, perhaps the most powerful, trend in educational policy in the last 20 years (Barber, 2004). In the 1980s, Prime Minister Margaret Thatcher's adoption of a standardized test-based accountability system in Great Britain provided a model for proponents of test-driven reform in the United States, Canada, Australia and other European nations (i.e., Germany, France). This type of test-driven system is often referred to as standards-based reform and still dominates much of the English-speaking western world. For example, in England, the trend is towards total accountability in education with the introduction of the National Curriculum in 1988 to the National Curriculum assessment system that was introduced in 1996 (Whetton, Twist, & Sainsbury, 2000). England currently measures progress against national standards when students reach the ages of 11, 14, and 16 (Holloway, 2003). League tables that summarize the performance of schools are published by local and national newspapers, attracting a

considerable amount of political and public attention. The underlying message conveyed to parents continues to be that they should be relatively satisfied with schools that improve their test performance from year to year, and begin to question the quality of instruction in those that have poor performance. Not surprisingly, accountability has become synonymous with external testing.

External testing enjoys a similar status in other western nations. In Ontario, Canada's largest province, testing is conducted under the direction of the Education Quality and Accountability Office (EQAO). Results are widely disseminated in a manner that invites comparisons across teachers, schools, and districts. Similar standardized testing programs operate throughout this country, garnering widespread media attention. In the United States, the federal No Child Left Behind Act (NCLB) requires every state to develop standards, standardized tests and accountability systems, and, by mandating the option for students to transfer from schools with low test performance to those with higher performance, NCLB promotes competition between schools (Hursh, 2005). Not surprising, the testing industry has rapidly expanded in the United States. In Victoria and New South Wales, Australia, standardized tests also perform a significant role in curriculum reform (Barnes, Clarke, & Stephens, 2000). This rapid expansion of standardized achievement testing and the impact of such measures on students and teachers, continues to provoke fierce debate.

The Consequences of High-Stakes Testing

Various studies have investigated the effects of mandated testing programs, particularly those with high-stakes attached to test results, over the last 20 years. The majority of these studies have addressed a variety of issues related to the effects on teaching and

learning; the strategies used to deliver instruction; the impact of the format of the test on classroom practices; the test preparation; and the psychological impacts of the test on both students and teachers, as well as on student learning in general (Abrams, 2004). Although the bulk of this research literature comes from the United States, the following section also draws from British, Canadian, and to a lesser extent, Australian literature in summarizing the positive and/or negative consequences associated with high-stakes testing. It should also be noted that the ensuing discussion does not necessarily represent a balanced view on this subject since most educationalist tend to dismiss the use of standardized testing as a lever for school improvement. Thus, the available research literature may be somewhat skewed to support this position.

Despite the previously noted limitation, there are pockets of research that suggest positive consequences for both students and teachers can accrue from high-stakes testing. Consider the following:

- In select American jurisdictions, the achievement of students in particular grade levels and subject areas increased substantially following the introduction of high-stakes testing (Roderick, Jacob, & Byrk, 2002).
- Research using anecdotal and secondary information suggested that students with disabilities have higher expectations placed on them and receive improved instruction when teachers are confronted with high-stakes assessment (Ysseldyke, Dennison, & Nelson, 2004).
- Teachers have made positive changes in their instructional and assessment practices as a direct result of receiving high-stakes test scores, particularly those

who received assistance from lead teachers and principals (Cizek, 2001; Herman, 2005).

- Teachers tend to increase their participation in staff development in tested subject areas and are more likely to take advantage of staff development programs linked to important standardized test measures (Earl & Torrance, 2000).

Although limited in scope, these findings suggest that particular jurisdictions have experienced some success with standardized testing. Interestingly, these pockets were primarily jurisdictions that accompanied their aggressive testing policy with significant investments in after school programming and/or included measures that contained more open-ended items designed to test a broader array of reading and writing skills.

In contrast to the previous section, extensive research has documented a number of important negative social consequences associated with high-stakes testing. Consider the following in relation to students:

- There is little evidence to support the proposition that high-stakes tests, including high school graduation exams, increase student achievement (Amrein & Berliner, 2003; Natriello & Pallas, 1999).
- There is also little evidence to suggest that the achievement gap will ever close as a result of test-based accountability. In fact, some research indicates that the gap is widening between low and high achieving students and/or that these tests continue to be an impediment to graduation for ethnic minorities (Boe & Shin, 2005; Gipps, 2003; McNeil, 2000; Scoppio, 2002; Valencia & Villarreal, 2003).

- Based on data from American states and Canadian provinces, high-stakes testing policies lead to increased dropout rates and decreased graduation rates (Amrein & Berliner, 2003; Hauser, 2001; Volante, in press).
- There is a trend toward grade retention for low performing students before pivotal testing years, apparently to ensure that these students are properly prepared and will pass these important tests (Amrein & Berliner, 2003; Hursh, 2005; Kornhaber, 2004).
- There are increased levels of stress and anxiety reported in response to standardized testing, particularly for students who do not achieve well or struggle in tested subject areas (Gipps, 2003; Lashway, 2001; Scott, 2007).

Collectively, these findings contradict the chief rationale underlying the use of high-stakes testing for accountability purposes – namely, that such measures spur improvements in the achievement levels of *all* students, and in doing so, promote school effectiveness.

Not surprisingly, research has also documented the negative impact of high-stakes testing on teachers. Consider the following:

- Because subjects such as art, music, physical education, social studies, and science are often not tested, teachers and administrators tend to focus less on these subjects as high-stakes testing dates approach (Bottrell & Ling, 2000; Burroughs, Groce, & Webeck, 2005; Earl, 1999; Graham & Neu, 2004).
- Teachers often employ “teaching to the test” techniques in preparation for high-stakes tests, limiting instruction to only those things that are sure to be tested, requiring students to spend hours memorizing facts, and drilling students on test-

taking strategies (Earl, Levin, Leithwood, Fullan, Watson, Torrance, Jantzi, Mascall, & Volante, 2003; Hursh, 2005; Volante, 2004).

- High-stakes tests, particularly those that rely heavily on multiple-choice questions that require students to recall facts from large bodies of knowledge, constrict curriculum and instruction to a focus on superficial content coverage, discouraging teachers from pursuing more challenging and rigorous study (Falk, 1996; McNeil, 2000).
- There are numerous instances of cheating by teachers and other school personnel in response to the pressures of high-stakes testing (Gipps, 2003; Lashway, 2001; Simner, 2000).
- High-stakes tests increase stress and decrease morale among teachers, particularly those working in inner-city schools (Black, 1994; Croft & Waltman, 2005; Leithwood, Steinbach, & Jantzi, 2000; Volante, in press).
- Lastly, the ranking that typically accompanies high-stakes testing tends to drive highly qualified and competent teachers out of schools serving the most vulnerable student populations (Berlak, 2001; Delphi, 1998; Hargreaves & Fink, 2006).

As in the previous paragraph, these findings suggest that high-stakes testing degrades, rather than spurs, improvements in classroom practice. Students and teachers both typically experience increased stress and anxiety, which raises psycho-social concerns around student engagement and job satisfaction. Overall, the research indicates that promoting student learning and system improvement via student achievement testing

remains a formidable task. Perhaps, it is time for policy-makers to seriously consider the alternatives.

Modifying Existing Assessment Systems

Proponents of standardized testing rightfully contend that educators and their allies are losing the political battles around testing by opposing the current tests without offering viable alternatives that allow the larger society to hold them accountable for the effects of their teaching (Covaleskie, 2002). Thus, the current challenge is to articulate sound assessment alternatives that may serve the dual function of improving classroom practice *and* student achievement, within an overarching accountability framework that circumvents the negative consequences of an over-reliance on low-level standardized tests. As the proceeding discussion will indicate, a great deal of experimentation is already occurring in select jurisdictions around the world. For example, classroom assessment (also referred to as curriculum-embedded assessment) is increasingly being utilized for accountability purposes in pockets of the United Kingdom, United States, Canada, and Australia (Wilson, 2004). Isolating the effective features within these alternative systems provides a general template for other educational leaders to consider. As the Australian Council for Educational Research (2002) argued, the challenge for policy-makers is to learn from past programs and to ensure that each decision made in designing an assessment system facilitates student learning.

Signs of change in national testing policy are slowly emerging, with trials to replace some testing programs and initiatives to promote formative assessment (Black & Wiliam, 2005). For example, both Scotland and Wales have decided to scrap most standardized testing through age 14, and an influential commission has recommended

that England replace testing of 11- and 14-year-olds with teacher evaluations (FairTest Examiner, 2007). England previously dropped the testing of 7-year-olds. Clearly, the tide is beginning to shift, and must do so, to support schooling that will encourage all students to construct, integrate, and apply their knowledge; to think critically and invent solutions to problems; and to respond creatively to issue that will confront them in the complex world of the 21st century (Falk, 1996). Large-scale assessment programs will have to adapt if they are to remain viable (Wilson, 1999).

A closer review of the literature suggests there are three main ways to modify existing assessment systems:

1. Redesign standardized achievement tests to reflect a greater emphasis on a broader range of skills, particularly those that tap critical and higher-order thinking;
2. Utilize a mixture of large-scale and classroom-based assessments in a synergistic accountability system;
3. Utilize only classroom-based assessments for accountability purposes.

Obviously, none of the previously noted approaches could claim complete superiority, particularly since each one as has important advantages and disadvantages. Deciding which option to pursue, depends on a variety of factors such as the available expertise of test designers, assessment literacy of classroom teachers, and resources to support continued professional development, to name only a few. Nevertheless, the status quo approach of utilizing high-stakes measures which emphasize low-level multiple-choice items is no longer tenable given the requirements of the knowledge economy.

On the surface, the first option may appear to be the most prudent given the variability in teachers' evaluative judgments. When the standardized test items and tasks change for each administration and are designed to push students beyond the recall of facts and algorithms to higher-order thinking and problem solving, the best strategy for the preparation of students is good teaching of all of the curriculum (Earl, 1999). Research in various American states also supports this contention, since standardized tests that require students to formulate and to provide written responses to test questions show an increased emphasis on writing and higher-level thinking skills (Abrams, 2004). Thus, well crafted tests, which include open-ended questions, may minimize and even eliminate "teaching to the test." Unfortunately, if teachers' own assessments do not play a part in summative testing, the summative will overrule and marginalize teachers' formative assessments unless the link between the two is carefully structured (Black, 1994; Black & Wiliam, 2005).

The second option involves the utilization of multiple forms of student assessment to inform accountability judgments. Recent research in the United States suggested that student achievement on standardized tests improved only in states using multiple student measures (Darling-Hammond, Rustique-Forrester & Pecheone, 2005). Indeed, the assessment systems of most of the highest-achieving nations in the world are a combination of centralized assessments that use mostly open-ended and essay questions and local assessment given by teachers which are factored into the final examination scores (Wood, Darling-Hammond, Neill, & Roschewski, 2007). The previous authors note that these local curriculum-embedded assessments tasks – which include research papers, applied science experiments, presentations of various kinds, and projects and

products that students construct – are mapped to standards for the subject and are selected because they represent critical skills, topics, and concepts. In general, redesign efforts around the world suggest that multiple forms of student assessment are critical to improving student achievement and promoting valid assessment (Herman, 1997; Koretz, Barron, Mitchell, & Keith, 1996; Wood et al., 2007). Clearly, balancing our assessment alternatives is a lofty goal, but one worthy of the effort.

Students also identify many of the positive features of alternative assessments commonly noted by researchers, particularly in relation to the impacts of these assessments on learning or its consequential validity (Sambell, McDowell, & Brown, 1997). In their study, the previous authors found that from the students' point of view, assessment has a positive effect on their learning and is fair when it:

- Relates to authentic tasks
- Represents reasonable demands
- Encourages students to apply knowledge to realistic contexts
- Emphasizes the need to develop a range of skills
- Is perceived to have long-term benefits
- Rewards genuine effort, rather than measuring 'luck'
- Rewards breadth and depth in learning
- Fosters student independence by making expectations and criteria clear
- Provides adequate feedback about students' progression
- Accurately measures complex skills and qualities, as opposed to an over-reliance on memory or regurgitation of facts

Sambell et al. (1997) concluded that the striking comparisons students drew between traditional and alternative assessment mechanisms suggest that an effective way to change student learning behavior is to alter the method of assessment. Essentially, curriculum-embedded assessment must factor into important decisions regarding student achievement.

The last option, complete abandonment of large-scale assessment, has considerable appeal for large numbers of educationalists. Nevertheless, this option is only viable if teachers' classroom assessments can sufficiently stand up to public scrutiny – a tall order given the relatively low levels of assessment literacy for large numbers of classroom teachers (Popham, 2004; Stiggins, 2002; Volante & Melahn, 2005). Teachers' idiosyncratic approaches to classroom assessment also make it impossible to compare student learning without common tasks and rating criteria. Rather, the available research suggests that there are distinct advantages to using a system that includes some standardization. For example, in Queensland Australia, the state's "New Basics" and "Rich Tasks" approach to standards and assessment offers extended, multi-disciplinary tasks that are developed centrally and used locally when teachers determine the time is right and they can be integrated with locally-oriented curriculum (Wood et al., 2007).

In truth, there is no 'royal road' to an assessment system that effectively serves both formative and summative functions (Black & Wiliam, 2005). It seems likely that the unique approach adopted by each country will require some standardization and flexibility for teachers to utilize curriculum-embedded assessment through a process of trial and error. A commitment to the underlying principles of *Learning-Focused*

Accountability provides the best prospect for reforming current top-heavy test-driven systems in a positive direction.

Learning-Focused Accountability

An over-emphasis on standardized testing orients accountability only to outcomes that can be conveniently measured and diverts attention from other purposes and goals for education such as good citizenship, social skills, technological competence, and preparation for employment (Ben Jaafar & Anderson, 2007). Even more disconcerting, particularly to the economic health of a nation, are assessment systems that fail to spur innovation and critical thought. Learning-focused accountability requires the effectiveness of an assessment system to be judged by teacher's ability to 'teach for transfer of learning' and student's ability to demonstrate that 'authentic learning' has actually taken place. This is in stark contrast to the common practice of looking at test score improvements for low-level literacy and numeracy skills that are easily malleable to manipulation. It is bitter irony that the pressure to improve school systems through standardized testing measures has actually stifled the use formative classroom assessment practices that have had the most significant impact on student learning and achievement (Black & Wiliam, 1998).

No government can decree an educated citizenship on the basis of improvement on a simplistic measure, since these test scores can go up with little or no change in learning (Earl, 1999; Smith & Fey, 2000). A preoccupation with authentic learning and critical/higher-order thinking should therefore become the benchmark upon which assessment systems should be judged. If we value critical and creative thought, it must be reflected in the assessment systems that influence teaching and learning. Failure to do so

stifles the creative capacities of teachers and students and prevents children from reaching their full potential. Clearly, test-based accountability models need to be replaced with more diverse and flexible assessment systems.

Conclusion

Although the research is not conclusive, the general trend suggests that current test-based accountability models which emphasize high-stakes testing have significant negative consequences for students and teachers. For the most part, this general approach has failed to deliver large-scale improvements in student performance or close the gap between high and low achievers in a broad range of contexts. Furthermore, when one considers the simplistic nature of many of these measures, the prospect of promoting higher-order or critical thinking skills is greatly diminished. The preceding discussion suggested that curriculum-embedded assessment within a learning focused accountability framework represents an opportunity to positively change national assessment systems. This type of approach should be held to the same level of scrutiny as test-based accountability models. Those in the education community will need to continuously research such models to answer the following critical questions:

- How accurate is curriculum-embedded assessment?
- How can assessment experts and professional development initiatives help teachers design and evaluate assessments with a sufficiently high degree of reliability?
- What kind of trade-offs between reliability and validity do curriculum-embedded assessment procedures produce?
- What consequences should policymakers attach to student performance to make alternative assessment systems work as intended?

- What are the costs of developing and maintaining alternative assessment systems?

Hopefully, the jurisdictions that are already experimenting with alternative systems will expand to allow more rigorous research on the previously noted questions. Certainly, the cracks in the test-based accountability armor present an opportunity to develop a healthy preoccupation with authentic teaching and learning and the assessment systems that measure such attributes.

This research is supported by a Social Science and Humanities Research Council (SSHRC) grant from the Canadian federal government.

References

- Abrams, L. M. (2004). *Teachers' Views on High-Stakes Testing: Implications for the Classroom*. Temple, AZ: Education Policy Studies Laboratory.
- Amrein, A. L., & Berliner, D. C. (2003). The testing divide: New research on the intended and unintended impact of high-stakes testing. *Peer Review*, 5(2), 31-32.
- Australian Council for Educational Research (2002). *Research Highlights*. Melbourne, Australia: Author.
- Barber, M. (2004). The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education*, 185(1), 7-38.
- Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: The engine of systematic curricular reform? *Journal of Curriculum Studies*, 32(5), 623-650.
- Ben Jaafar, S., & Anderson, S. (2007). Policy trends and tensions in accountability for educational management and services in Canada. *Alberta Journal of Educational Research*, 53(2), 205-225.
- Berlak, H. (2001). *Academic achievement, race, and reform: Six essays on understanding assessment policy, standardized achievement tests, and anti-racist alternatives*. California: United States Department of Education.
- Black, P. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational Evaluation and Policy Analysis*, 16(2), 191-203.
- 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.
- Boe, E. E., & Shin, S. (2005). Is the United States losing the international horse race in academic achievement? *Phi Delta Kappan*, 86(9), 688-695.
- Bottrell, C., & Ling, L. (2000). *The Future: Optimism or Ossification*. Paper presented at the Annual Meeting of the Australian Association for Research in Education, Sydney, Australia.
- Burroughs, S., Groce, E., & Webeck, M. (2005). Social studies education in the age of testing and accountability. *Educational Measurement: Issues and Practice*, 24(3), 13-20.
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27.
- Covaleskie, J. F. (2002). Two cheers for standardized testing. *International Electronic Journal for Leadership in Learning*, 6(2). Retrieved August 6, 2006, from <http://www.ucalgary.ca/~iejll/volume6/covaleskie.html>
- Croft, M., & Waltman, K. (2005 April). *The Impact of School-Level Accountability on Local Test Preparation Practices*. Paper presented at the National Council on Measurement in Education, Montreal, Quebec.
- Darling-Hammond, L., Rustique-Forrester, E., & Pecheone, R. L. (2005). *Multiple Measures Approaches to High School Graduation: A Review of State Student Assessment Policies*. Stanford, CA: School Redesign Network. Retrieved August 6, 2006, from http://www.schoolredesign.net/srn/mm/pdf/multiple_measures.pdf.

- Delphi, K. (1998). Shopping for schools. *Orbit*, 29(1), 29-33.
- Earl, L. M. (1999). Assessment and accountability in education: Improvement or surveillance. *Education Canada*, 39(3), 4-6.
- Earl, L., Levin, B., Leithwood, K., Fullan, M., Watson, N., Torrance, N., Jantzi, D., Mascall, B., & Volante, L. (2003). *England's National Literacy and Numeracy Strategies: Final Report of the External Evaluation of the Implementation of the Strategies*. England: Department for Education and Skills.
- Earl, L., & Torrance, N. (2000). Embedding accountability and improvement into large-scale assessment: What difference does it make? *Peabody Journal of Education*, 75(4), 114-41.
- FairTest Examiner (January, 2007). *Wales Drops Most Standardized Testing*. Retrieved September 20, 2007 from <http://www.fairtest.org/examarts/2007%20January/Wales.html>.
- Falk, B. (1996). *Issues in Designing a Learner-Centered Assessment System in New York State: Balancing Reliability with Flexibility, Authenticity, and Consequential Validity*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.
- Gipps, C. V. (2003 April). *Educational Accountability in England: The Role of Assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Graham, C., & Neu, D. (2004). Standardized testing and the construction of governable persons. *Journal of Curriculum Studies*, 36(3), 295-319.

- Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA: Rand.
- Hargreaves, A., & Fink, D. (2006). The ripple effect. *Educational Leadership*, 63(8), 16-21.
- Hauser, R. (2001). Should we end social promotion? Truth and consequences. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards and raising barriers? Inequality and high-stakes testing in public education* (pp. 151-178). New York: Century Foundation.
- Herman, J. L. (1997). *Large-Scale Assessment in Support of School Reform: Lessons in the Search for Alternative Measures*. Los Angeles, CA: Center for the Study of Evaluation.
- Holloway, J. H. (2003). A global perspective on student accountability. *Educational Leadership*, 60(5), 74-76.
- Hursh, D. (2005). The growth of high-stakes testing in the USA: Accountability, markets and the decline of educational equality. *British Educational Research Journal*, 31(5), 605-622.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31-41.
- Koretz, D., Barron, S., Mitchell, K., & Keith, S. (1996). Perceived effects of the Kentucky instructional results information system (KIRIS). Santa Monica, CA: Rand.
- Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy*, 18(1), 45-70.

- Lashway, L. (2001). *The New Standards and Accountability: Will Rewards and Sanctions Motivate America's Schools to Peak Performance?* Washington, DC: ERIC Clearinghouse on Educational Management.
- Leithwood, K., Steinback, R., & Jantzi, D. (2000 April). *Identifying and Explaining the Consequences for Schools of External Accountability Initiatives or "What in the World Did You Think I Was Doing before You Came Along?"* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- McNeil, L. (2000). *Contradictions of School Reform: Educational Costs of Standardized Testing*. New York: Routledge.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18.
- Natriello, G., & Pallas, A. M. (1999). *The Development and Impact of High Stakes Testing*. New York: United States Department of Education.
- Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 16-18.
- Popham, W. J. (2004). All about accountability / Why assessment illiteracy is professional suicide. *Educational Leadership*, 62(1), 82-83.
- Pringle, R. M., & Martin, S. C. (2005). The potential impacts of upcoming high-stakes testing on the teaching of science in elementary classrooms. *Research in Science Education*, 35, 347-361.
- Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: issues and Practice*, 17(2), 13-16.

- Roderick, M., Jacob, B. A., & Bryk, A. S. (2002). The impact of high-stakes testing in Chicago on student achievement in promotional gate grades. *Educational Evaluation and Policy Analysis, 24*(4), 333-357.
- Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*(4), 349-371.
- Scoppio, G. (2002). Common trends of standardization, accountability, devolution and choice in the educational policies on England, U.K., California, U.S.A., and Ontario, Canada. *Current Issues in Comparative Education, 2*(2), 130-141.
- Scott, C. (2007). Stakeholder perceptions of test impact. *Assessment in Education, 14*(1), 27-49.
- Simner, M. L. (2000). *A joint position statement by the Canadian Psychological Association and the Canadian Association of School Psychologist on the Canadian press coverage of the province-wide achievement test results*. Retrieved July 7, 2005 from http://www.cpa.ca/documents/joint_position.html.
- Smith, M. L., & Fey, P. (2000). Validity and accountability of high-stakes testing. *Journal of Teacher Education, 51*(5), 334-344.
- Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-65.
- Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement and Practice, 17*(2), 20-23.
- Valencia, R. R., & Villarreal, B. J. (2003). Improving students' reading performance via standards-based reform: A critique. *The Reading Teacher, 56*(7), 612-621.

- Volante, L. (in press). Equity in multicultural student assessment. *Journal of Educational Thought*, 42(1).
- Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, 35. Retrieved December 13, 2004 from <http://www.umanitoba.ca/publications/cjeap/articles/volante.html>.
- Volante, L., & Melahn, C. (2005). Promoting assessment literacy in teachers: Lessons from the Hawaii School Assessment Liaison Program. *Pacific Educational Research Journal*, 13, 19-34.
- Whetton, C., Twist, E., & Sainsbury, M. (2000 April). *National Tests and Target Setting: Maintaining Consistent Standards*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Wilson, R. J. (1999). Aspects of validity in large-scale programs of student assessment. *Alberta Journal of Educational Research*, 45(1), 333-340.
- 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.
- Wood, G. H., Darling-Hammond, L., Neil, M., & Roschewski, P. (2007). *Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills*. Amesville, Ohio: Forum for Education and Democracy.

Ysseldyke, J., Dennison, A., & Nelson, R. (2004). *Large-Scale Assessment and Accountability Systems: Positive Consequences for Students with Disabilities*. Minneapolis, MN: National Center on Educational Outcomes.