A Balanced Approach to High-Stakes Achievement Testing:An Analysis of the Literature With Policy Implications

John M. Burger

John.Burger@gov.ab.ca
Alberta Learning

and

Monte Krueger

Monte.Krueger@gov.ab.ca
Alberta Learning

Abstract

This article examines the intended and unintended impacts of large-scale, high-stakes achievement tests on teachers and students through analysis of the arguments put forth by testing advocates and critics. A key objective is to inform both the evolving dialogue around testing and assessment policy development. Attention is given to: (1) the role of values and beliefs as determiners of one's approach to testing, (2) methodological issues, (3) the impacts of testing on teacher and student behaviours, and (4) accountability and political implications of testing. A key conclusion of the article is that many of the arguments for and against testing concern the issues of fairness and usefulness of testing for teachers and students, juxtaposed with the public's and government's need to know how well schools are performing. Reconciling the two sides in the interest of advancing assessment policy is a matter of building the case that a balanced approach to testing exists. Such an approach must give appropriate attention to the multiple functions of classroom assessment relative to the functions of high-stakes achievement testing. Summative, formative, criterion-referenced, and norm-referenced testing models all have specific attributes and benefits that fit varying testing needs in given situations. However, finding the right balance in their applications and usages is the key to building a fair and integrated model of classroom assessment.

Introduction

Accountability for students is related in important ways to accountability of educators, schools and school districts. Indeed, the use of tests for accountability of educators, schools and school districts has significant consequences for individual students. Such indirect effects of large-scale assessment are worth studying in their own right. (Heubert & Hauser, 1999, p. 2)

"The use of large-scale achievement tests as instruments of educational policy is growing" (Heubert & Hauser, 1999, p. 1). There is considerable debate as to whether or not this is as it should be. This article discusses the impact of large-scale achievement tests on students and teachers and considers the debate from the perspective of advocates for and critics of high-stakes testing in order to fully understand the issues involved. The state of the current discourse regarding this issue is fueled by rhetoric and emotion and is heavily weighted with literature pertaining to the United States, where testing is a hot political topic. The purpose of this work, then, is to inform a fair and unbiased dialogue regarding provincial, national, and international high-stakes testing as a precursor to surveying ideas on how to optimize the use of high-stakes test data specifically, and classroom assessment data more generally with specific reference to the Alberta provincial achievement testing model (Alberta Learning, 2003).

The term "high-stakes achievement tests" is defined as including either norm-referenced or criterion-referenced tests mandated by the state where the aggregated results are used for summative purposes such as judging the effectiveness of programs or schools. It must be stressed, however, that these summative applications do not necessarily preclude the use of the test data in a formative manner when reviewing individual student achievement.

In this paper, we discuss: (1) the role of values and beliefs as factors that shape perspectives on testing; (2) methodological issues; (3) the impacts of testing on both teacher and student behaviour; and (4) the political/accountability implications of testing.

Values and Beliefs

On one side of the debate over testing as policy instruments are proponents of accountability who believe the quality of learning is enhanced by measuring performance of educational programs or policies against student outcomes. "The government of Prime Minister Tony Blair has based its approach on [the] belief that educational change is doomed without equal measures of government pressure and comprehensive support for those on the ground" (Schofield, 2001, p. 6). Likewise, the Ontario Ministry of Education's Task Force on Effective Schools claimed in its report that "effective educational change always needs a blend of pressure and support" (2000, p. 60). Thus, a major assumption underlying and directing the action of the policy actors that advocate high-stakes achievement testing is the need to make decisions about program or school effectiveness based on well-balanced data and rich critical analyses. In this light, high-stakes achievement tests decrease the level of subjectivity in policy decision-making while helping in the evaluation of the efficacy of learning programs and systems. Such measures support decisions about the best way to improve curriculum, establish which programs are working and which are not, and about how best to implement changes that will both inform and increase the "public's confidence in schools" (Heubert & Hauser, 1999, p. 1).

Supplemental to the policy/program impact of standardized assessment, there is the need to consider the usefulness of high-stakes achievement test results at the individual student/parent/teacher levels. In this context, parents often have three information needs: how their child is doing in relationship to the standards inherent in the program of study; how their child is doing relative to the child's peers; and how their child's school is doing in relation to other schools of choice.

On the other side of the debate are those who oppose high-stakes achievement tests and claim that such exams increase the risk of education failure, hold teachers responsible for results with inequitable resources, narrow and distort curriculum, are not relevant indicators of student's achievement, and solidify class and racial disparities. Most arguments against high-stakes achievement tests are based on objections to policy-maker's desires to quantify and objectify a system that opponents feel is subjective and not quantifiable, or that single tests provide a minimalist perspective of system results. Some, such as Kohn (2000), even claim that the system would work best if student performances were evaluated on contextual frameworks that employed continuous evaluation and "portfolios" rather than tests. In this light, it becomes apparent that the two sides in the debate are distinguished by methodological, political, or value based differences. Indeed, Runte (1998) contends that high-stakes achievement testing is leading to the "proletarianization" of teachers, thus reflecting the embedded ideological underpinnings that sometimes demarcate the sides. Likewise, in regard to the English Education Reform Act that advocated high-stakes achievement testing, Simon (1993, p. 32) quoted J. Marlowe, who wrote in the May 26, 1991, Observer that "the legitimization of the whole [testing] policy rests precisely on the image of schooling that the press, some … industrialists, ideologists, and politicians have created."

The unfortunate result of this variation in perspective is a debate "that has too often been characterized by misrepresentations and jargon" (Rothman, 1995, p. xiv). Related to this, the sides seem to be divided based on their values. Core values in a policy framework are the deeply held beliefs that certain groups, or advocacy coalitions, "clump" around. In other words, groups are bound together by shared belief systems, and these beliefs largely determine one's participation and position in a debate. The importance here is that embedded value assumptions help to explain some of the discrepancy discussed below in the debate over high-stakes achievement testing. Specifically, differences in values will largely explain why one group cites an unintended impact as negative whereas the other views the same impact positively.

In order to transcend the ideological dichotomy one needs to evaluate, definitively and objectively, the arguments being made for and against high-stakes achievement tests. More critically, it is necessary and worthwhile to try and isolate the intended and unintended ripple effects that such testing may create, because as the American Educational Research Association (AERA, 2000, p. 1) points out, "although policy makers generally institute [high-stakes achievement] tests with the good intention of improving education, they need to carefully evaluate the tests' potential to create positive impacts or to cause serious harm." Therefore, the focus of this article is to isolate both the positive and negative, intended and unintended impacts of high-stakes achievement tests, with specific concern for the particular attributes of the Alberta testing model. Doing so will hopefully elicit a more informed and balanced dialogue that will contribute to a richer understanding of testing implications and issues.

Methodological Issues

As stated earlier, proponents of high-stakes achievement testing claim that the tests provide objective measures by which to evaluate, critique, and improve education systems. In short, high-stakes achievement tests are thought to be an integral facet of the "feedback loop" that inputs new, trustworthy, and accurate information into the policy network, so that programs and initiatives may be monitored and improved as part of a mutable process. In this way, problems are hopefully dealt with before they "boil over" so that teachers and students are not subjected to sudden, unsubstantiated, or misdirected system reforms. Additionally, proponents like Lewis (2000, p. 3) believe that high-stakes achievement testing leads to efficacy in teaching and learning, as the tests "encourage teachers and students to get serious about learning," by providing clear objectives or targets at which to aim. This in turn enables clear and focused curriculum delivery.

On the surface, the merit of the aforementioned arguments put forth by advocates of high-stakes achievement testing seems obvious. However, they are based on the assumption that the tests are objective: an assumption that opponents (Kohn, 2001, p. 4) are quick to challenge by asking, "Is objectivity really a desirable or realistic [condition]?" Essentially, the methodological question is centred on the reliability and validity of high-stakes achievement testing, because "while some tests may validly predict future performances of groups, critics of testing argue that they are often inaccurate predictors of individual performance" (Worthen & Spandel, 1991, p. 65). This claim concerns students who typically perform well academically but may attain a poor standing on the examination because they had a "bad day" or undervalued the significance of the tests. An acceptance of this reality cannot in itself be considered a flaw in high-stakes achievement testing. However, if the stress or attitudes associated with writing those tests is sufficient to prompt less than indicative results from many students or specific subgroups of students on an ongoing basis, there may be a systemic validity issue associated with the tests.

It is important to realize that not all aspects of student learning may be measured in an objective fashion by paper–and-pencil tests. Some aspects of learning are better suited to performance-based measures rather than discrete, selected-response indicators. Reading comprehension is one such area. If a student is given a passage to read and then asked specific questions about it, the ability to answer the questions does not necessarily reflect the student's level of comprehension. Aspects of the passage may have been internalized and comprehended in a way that would not be immediately apparent by asking one set of questions. Nonetheless, with high-stakes achievement testing, it is possible that decisions will be made about the students' performance on the basis of the test results, although the tests results are not sufficient to adequately assess the students' knowledge or skills, and to support the decisions being made. If there is a low level of consistency to the testing methodology or the administration thereof, the information in the feedback loop would likewise be skewed, and the entire process itself may be subject to higher–than-acceptable levels of error. Hence, if the underlying assumptions of objectivity, validity, and reliability cannot be demonstrated, the generalizations made on the basis of the tests would be suspect.

This is not to say the impacts of high-stakes achievement tests are necessarily problematic. To the contrary, Cizek (2001, p. 24) isolates two major areas where positive "spin-off" effects are

evident. First, the advent of high-stakes achievement testing may have spawned a corresponding positive increase in educator's knowledge of testing and of testing issues: "Increasingly, teachers can tell you the difference between a norm-referenced and a criterion-referenced test; they can recognize, use, or develop a high-quality rubric; they can tell you how their state's writing test is scored, and so on." In this case, necessity has been the mother of invention. A likely effect of this, Cizek notes, could be a more efficient use of professional development time: "Driven by the demands of high-stakes tests, the press toward professional development that helps educators hone their teaching skills and content area expertise is clear" (p. 23). The idea seems to be that time set aside for teachers' professional development is increasingly focused on curriculum-relevant, results-oriented materials. This could account in some way for the perceived rise in knowledge about psychometric issues and their relationship to effective pedagogy amongst teachers in general.

Second, the collection and use of measurement information have advanced. "Obtaining information about test performance, graduation rates, per-pupil spending, staffing, finance, and facilities is, in most states, now just a mouse click away" (Cizek, 2001, p. 24) [1]. And importantly, the data are not only available; they may be collected and managed in a conscientious, accurate, and holistic manner reflecting an increased knowledge of statistical methodologies and meaning of the interrelationships between a range of important variables affecting student learning. However, these advances may come with some related risks. As knowledge of the testing procedures increases, so too does knowledge of assessment techniques. Thus, there is a danger that some teachers who have access to test items or test blueprints may "teach to the tests" or focus their teaching on only those knowledge or skill domains that they know will be tested to the exclusion of curriculum content not tested. To conclude this section, the issues associated with the psychometric implications of high-stakes achievement testing are summarized in Table 1.

Table 1: Methodological Issues

(+) Positive Implications	Main Issues: Speak to Fairness and Objectivity in Testing	(-) Negative Implications
If tests are reliable, results may help to inform curriculum and teachers' pedagogy	Reliability	Poor reliability limits test usefulness for judging programs or levels of student achievement
Valid tests help inform the policy, administrative and classroom decision-making processes	Validity (construct and predictive)	If not valid, test results can misinform policy, administrative and classroom decision-making processes
More efficient professional development hones teaching and content area expertise	Teacher knowledge of methodological issues is increasing	Teachers use testing knowledge to manipulate the test results

Impacts on Teacher and Student Behaviours

Some teachers may internalize test results, and translate poor results into feelings of guilt or shame. Often, these feelings are based on a belief that they have failed to appropriately or adequately prepare their students for the test. As a means of rationalizing, the test may be blamed and teachers may begin to question the usefulness and necessity of testing. These views can be picked up by the students, who in turn develop their own feelings of anxiety. Thus, the whole situation may become somewhat circular; the simple fact is, both groups may be affected by high-stakes achievement testing, and that fact deserves attention.

The emotional impacts on students can be broken down into three broad categories:

- Apathy or anxiety about tests
- Hyper-motivation to succeed, or lack of will to give genuine effort
- Increased use of inappropriate strategies

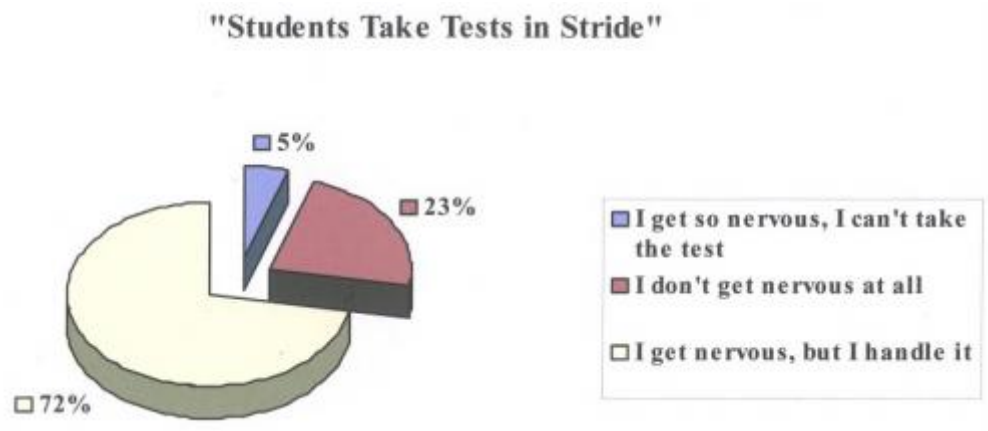
Paris, Lawton, Turner, and Roth (1991, p. 15) found that younger children are more trusting of their teachers, and are generally willing to accept that high-stakes achievement tests are accurate and useful measures of their achievements. Conversely, older students "feel less informed about the purposes and uses of high-stakes achievement tests, perhaps because adolescents expect more information about the value of tasks given to them in school." Paris et al. also found that as the age of students increased, so too did student skepticism that high-test scores were reflective of effective teaching. So, with each successive year of high-stakes achievement testing, the concurrent validity of the tests with effective teaching practices seems to decrease. This point becomes very problematic. Older students most likely realize that the tests are being used to make decisions about students, and that the tests may also be used to rank students. As Cizek (2001, p. 21) states, "there is simply no way to escape making decisions about students. These decisions, by definition, create categories." Admittedly, decisions about students would be made with or without high-stakes achievement tests, as teachers make decisions about students routinely.

However, what is important here is the fact that students may develop associations between high-stakes achievement tests and decisions being made about their lives. When this is coupled with an increased level of skepticism about the testing process, high-stakes achievement tests may contribute to barriers to school completion for some students. If students believe that the tests are not accurate representations of their abilities, or that the tests represent an impossible level of achievement, or that the tests are being used to judge or label them, the students may become disillusioned and simply exit from a system they feel is flawed. Indeed, Paris et al. (1991, p. 15) noted that "reduced effort thus reflects a developing awareness that tests scores may become the basis for comparative social judgments." Using the same logic, even if high-stakes achievement tests do not necessarily contribute to barriers to completion, they may have the "off-shoot effect" of influencing student motivation to succeed or learn. "Decreasing motivation may reflect an attempt by students who do not excel on the tests to protect their own self-esteem because they can always claim a lack of effort as an excuse." When this happens, some literature suggests, students are more likely to employ inappropriate strategies for test writing and preparation. Such things include filling in test bubbles to draw Christmas trees, or surfboards, or simply not

preparing for the exams at all. On the whole, a balance is likely achieved between the need to make decisions about students and the impact decision-making has. But, at the individual level, serious problems can result for a proportion of students, unless steps are taken to involve the students in the testing process to make it understandable and meaningful for them (Stiggins, 2001).

Public Agenda's Reality Check 2002, a survey of 600 middle and high-school students in the United States, seems to indicate that the majority of students are not "unnerved by what is being asked of them. Although some educators have raised concerns about whether increased standardized testing is placing undue pressure on students, Reality Check picks up little evidence of strain." Almost all students surveyed maintain they take the tests seriously, and 56% say they take them "very seriously." Only 23% state they do not get nervous at all, and 73% say they do get nervous, but that it is manageable stress "they can handle." Only 5% of the 600 students say "they get so nervous" they cannot take the test. The results were reported on Education Week's website (http://www.edweek.org/ew/newstory.cfm?slug=25realitycheck.h21) with a ±4% margin of error (see Fig. 1).

Figure 1. Student attitudes towards Testing
Source: Public Agenda (2002).



For those students who do value testing, a different set of responses can develop. With a belief in the importance of the testing may come an increased desire to succeed. The resulting internalized pressure could subject students to a high degree of stress, which can be manifested with positive or negative results. Dounay (2000, p. 4) reports that "stories of increasing numbers of children suffering from sleep disorders and other stress related maladies have appeared in the press in the past few years." She speculates that this stress could be directly related to summative high-stakes achievement testing, although admittedly, little attention has been given to the subject. Roderick and Engel's (2001, p. 219) qualitative study of the impact of high-stakes achievement testing on students in the Chicago public schools concluded that "creating incentives for low-achieving students through goals that provide an opportunity for feedback, a tangible reward, and a way to construct meaning regarding learning may have a positive impact on their motivation and effort in school." However, Roderick and Engel also identified "a group of students who did not respond to the pass incentive [linked to high-stakes test results] with significant effort." They also noted that "Students with the lowest skills were the least likely to respond positively." Casas

and Meaghan (2001, p. 150), considering Paris et al.'s 1991 study, concluded that "test anxiety is a chronic problem and correlates with damage to the self-concept."

Test anxiety may be even greater in younger children, and formal, summative testing may begin for some children at age six, [2] "despite the fact that almost all experts in early childhood education condemn early summative or norm-referenced testing" (Kohn, 2000, p. 2). When this is coupled with the realization that "[U.S.] children are tested or are facing proposals for increased testing to an extent that is unprecedented in our history and unparalleled anywhere else in the world" (p. 2), one begins to see why concerns are concentrated in the U.S. literature. It is estimated that the equivalent of "20 million school days are spent each year by American students just taking tests (and perhaps 10–20 times that many days are spent in preparation for the tests)" (Paris et al., 1991, p. 13). At present, it is unclear what the equivalent figures for Canada would be, but the case can be made that if the time spent in summative, norm-referenced testing is not a small proportion of the time allocated to formative, criterion-referenced testing, then students may be subjected to an ill-considered approach to classroom assessment.

The Alberta model of testing and its guiding principles, in contrast to many American programs, employs very well constructed criterion-referenced exams, blueprinted on a highly standard curriculum. Also, test results are reported as the percentage of students achieving the desired standards rather than the more typical descriptive statistics. Nonetheless, when the above figures are combined with the findings of Roderick and Engel (2001), Paris et al. (1991), and Dounay (2000), the obvious corollary is to suggest that North American children may be facing increased stress-loads owing to increased high-stakes achievement testing with mixed benefits dependent on a wide range of mediating variables, such as the extent to which test results are used formatively to inform students and their parents.

Stress over testing, however, is not limited to students. Smith (1991, p. 9) studied the effects of high-stakes achievement testing on teachers and developed specific categories of negative impacts. First, she determined that feelings of embarrassment, guilt, anger, and shame might be solicited in teachers when the results of low scores are made public. Obviously, if scores are published, teachers may be able to draw comparisons between their classes and others. If their class's test scores are low, they might feel they failed to adequately prepare the students. The perceived failures might in turn create internalized anxieties in teachers that adversely affect their performance. On the opposite end of the scale, if their class's scores were high, they might feel an increased pressure to duplicate the results or might come to rely on the effect of high socioeconomic status (SES) variables on student results. Unfortunately, whether anxiety is generated by low or high results, Smith (1991, p. 9) notes that feelings of anxiety often develop a determination in teachers to "do what is necessary to avoid such feelings in the future," such as avoiding teaching class levels that are tested directly. Smith's (1991) research suggests there is a need to position high-stakes achievement test results in ways that empower and inform teachers' understanding of their students' achievement levels and needs. In their study of high-stakes test impacts in the Chicago public schools, Roderick and Engel (2001, p. 221) concluded that "the instructional response of teachers, that is, whether they provide the support that students need to address learning problems, is critical in determining student [achievement] outcomes."

Saunders and Rivers (1996, p. 7) further examined the role of the teacher in student achievement, using the Tennessee Value-Added Assessment System (TVAAS), which was designed to "determine the influence of individual teachers on the rate of academic growth for student populations." Saunders and Rivers found that there was a profound cumulative effect of proficient teachers on student achievement. Earlier research had determined that "within grade levels, the single most dominant factor affecting student gain [was] teacher effect," and the subsequent study showed "residual effects [both positive and negative] of both very effective and ineffective teachers were measurable two years later, regardless of the effectiveness of teachers in later grades." For the purposes of this article, two main ideas may be derived from Saunders and Rivers' findings. First, it is obvious that the role of the teacher is a critically important factor influencing student achievement, and thus test data are an important source of information for teacher self-assessment. Second, Saunders and Rivers' work echoes the earlier claim that there is a need for a balance between pressure and support when utilizing high-stakes achievement testing results. In other words, that work shows that adequate support should be given to teachers in order to maximize their effectiveness in the classroom. In this light, accountability and data-rich critical reflection can be seen as tools leading towards continuous improvement.

As noted earlier, an emphasis on testing and test data may raise the risk that some teachers will "narrow" the curriculum and "teach to the test." If teachers focus on the material that will appear on the exam to the exclusion of other curriculum items, the subsequent conclusions drawn on the basis of the results will have limited predictive validity. In other words, teachers, parents, and others "will not be able to make accurate inferences about the levels of mastery that students have achieved with respect to a body of knowledge or a set of skills" (Popham, 2001, p. 2). Hess and Brigham (2000, p. 28) noted this fact and stated, "statewide assessments are not meant to suggest that only what is on the test is important, but many schools have interpreted them this way." In other words, areas that are "critically important" may not be "exclusively important," and thus, unless items are specifically included on the tests or teachers are otherwise accountable for teaching the whole curriculum, they may not be taught in the classroom. In addition to this, Hess and Brigham (2000, p. 29) note, "[summative, norm-referenced] testing generally emphasizes content knowledge rather than higher thinking or development skills," and the latter are also important life/career skills, which are not as easily evaluated by quantitative measures. In this respect, high-stakes, summative achievement tests may have the capacity to make teachers, administrators, and policy-makers decide which curricular knowledge and skills are more valued. Conversely, Popham (2001, p. 5) notes that if teachers and administrators can be convinced that "item-teaching" or teaching to the test is an inappropriate strategy for test preparation, then (as Saunders and Rivers have noted) improvements in test scores may result from properly conceived curriculum teaching. Therefore, rather than focus exclusively on the accountability aspect of testing, it is also important to view tests as supporting curriculum standards to help teachers develop the necessary content competencies in students that will lead to curriculum mastery and thus higher test scores.

The stress factor that teachers cite as a negative implication associated with high-stakes achievement testing is, paradoxically, a factor proponents see as positive. As stated above, many teachers experience stress when the results are made public. Those in favour of high-stakes achievement testing claim test stress is an acceptable trade-off for the greater availability of information and choice for parents and students (Cizek, 2001, p. 24). It is not unreasonable to

think that owing to the increased availability of information and a wider array of choices, parents may be prompted to become more involved in their children's education. Thus, there are those who believe that teachers should be made more accountable, and that parents and students should not only have the right to scrutinize educators and educational institutions, but that there are clear benefits in doing so.

Undeniably, policy-makers use high-stakes achievement testing as a policy instrument. The distinct policy purpose it serves, however, has varying interpretations. Heubert and Hauser (1999, p. 33) note that some feel the main usefulness of testing is formative, whereby it aids "instructional decisions about individual students," whereas others, they note, contend it is summative and "provides information about the status of the educational system." A third view of testing is that it can be a policy tool used to motivate change by "shaking people up" or "embarrassing them into change" (p. 34). It is important to note that such policies are not the norm, and are typically aimed at parents and educators, not students. Nonetheless, when policy-makers view testing as motivation for change, it can pit sides against each other, especially if the resulting policies are premised on finding fault. The result is a growing debate over the benefits versus disadvantages of high-stakes achievement testing.

Cizek (2001) argues that resistance of some in the teaching profession to the accountability movement (of which high-stakes achievement testing is at the centre) is nothing more than an attempt to maintain a flawed homeostasis. He claims, "the rationale is rational when there is a choice between being accountable for performance or maintaining a status quo without accountability" (p. 22). This seems to be an unrealistic choice in an era of increasing accountability.

To conclude, the main aspects considered in this section are summarized in Table 2.

Table 2: Impacts on Teacher and Student Behaviours

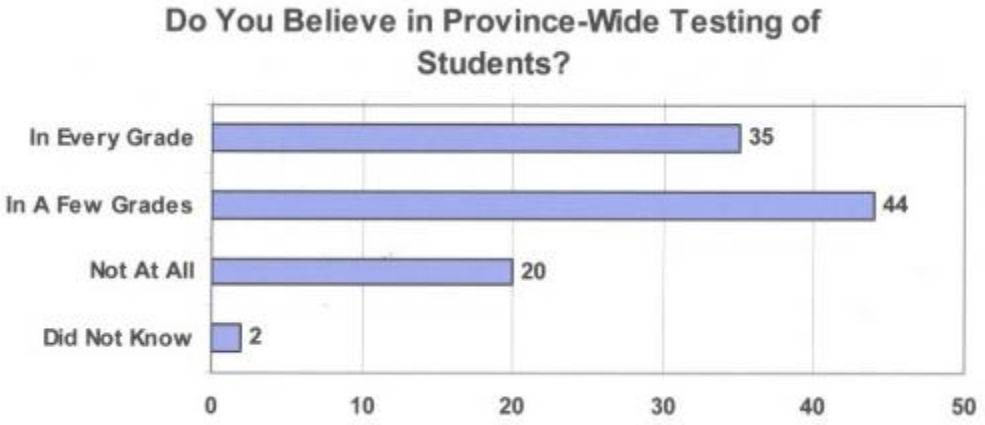
(+) Positive Implications	Main Issues	(-) Negative Implications
Teachers may use test results to inform their pedagogy	Teacher response to testing	Teachers may teach the test to the exclusion of non-tested curriculum
Test might motivate student achievement	Student response to testing	Tests might inhibit learning and depress motivation
Parents may have a better understanding of their child's level of achievement relative to the curriculum or peers	Parent response to testing	Information may not be communicated clearly to parents, leaving them to question the value of high-stakes achievement testing

Accountability and Political Implications

The following section examines more closely the relationships between accountability and the political implications of testing. However, two important facts should be highlighted regarding

the political nature of the debate. First, the literature seems to suggest that professionals are evenly divided into camps for and against testing. This is a result of the fact that the dialogue has become highly politicized and reflects the strongly held opinions of vocal opponents and proponents. Yet to juxtapose this debate within the profession to the broader public view, we consider the results of a recent survey of 785 Canadian adults conducted by Compas Inc. for the National Post/CanWest. This survey suggests that the general public is quite firm in their beliefs about the usefulness of testing. Thirty-five percent of survey respondents believed in province-wide testing for every grade and 44% believed in testing for a few grades, whereas only 20% did not believe in testing at all (see Fig. 2). The second important fact worth noting is that the public political discourse is largely generic and does not differentiate between the types of testing programs and their usages. Conversely, the professional/academic debate is much more specific and often does distinguish between summative, formative, norm-referenced, or criterion-referenced testing regimes.

Figure 2. Public opinion of the usefulness of testing
Source: Compas Inc. (2001).



Worthen and Spandel (1991, p. 67) claim, "perhaps the most serious indictment aimed at both norm-referenced and minimum competency [criterion-referenced] tests is that they may be biased against ethnic and cultural minority children." At the root of such strong claims is the notion that norm-referenced test scores really only give a reflection of socio-economic status, and perpetuate cultural capital. Kohn (2000, p. 7), citing a 1992 NAEP study of math scores, claims "the main thing [the tests] tell us, is how big the students' houses are." The NAEP chose four variables that reflected the students' socioeconomic backgrounds (number of parents living at home, parental levels of education, type of community, and state poverty rate), which they used as predictors of achievement. These four non-instructional-related indicators explained 89% of the variance. Kohn then furthers his argument by citing Roger's (1997) study conducted in Edmonton, and claims, "socio-economic status was by far the strongest predictor, accounting for the vast majority of variability in grade three and grade six achievement scores"(p. 68). Worthen and Spandel (1991, p. 67), who contend that "most published tests favor economically and socially advantaged children over their counterparts from lower socio-economic families," echo this point. These arguments tend to support Coleman et al.'s (1966) report that pointed to the strong relationship between external variables and student achievement, but ignore the potential

strategic value of high-stakes achievement testing for informing a range of decision-makers, from policy-makers to classroom teachers and students and parents, who are in a position to influence important school-related variables associated with student achievement.

A recent OECD Programme for International Student Assessment (OECD, 2000, p. 5) survey showed that Alberta students, compared to similar students in other jurisdictions, had higher average scores in all domains tested, regardless of their SES. "Alberta had generally the highest achievement scores across all levels of family socio-economic background, yet had greater variation in scores across socio-economic groups." This means that jurisdictions with some form of a high-stakes achievement testing model, such as Alberta, do better than those without, but supports the idea that external variables still play a role. Student achievement is always, to some extent, influenced by factors that are beyond the control of the school, such as SES.

McNeil (2000) cites Haney's (1999) report wherein the latter, beginning in 1978, tracked students from Grade 9 to graduation. In 1978, approximately 60% of black and Latino students graduated, compared to roughly 75% of white students. "By 1990, after four years of Perot-era standardization reforms, [3] graduation rates for blacks, Latinos and whites had all dropped" (McNeil, 2000, p. 732). By 1999, the rate for white students had rebounded to its 1978 level of 75%, but for blacks and Latino students it remained at around 50%. Therefore, the reforms seemingly had the effect of widening the gap between minorities and white students. In light of this fact, it again seems pertinent to consider the concurrent validity of the tests or the strength of the relationship between test results and students' actual levels of learning. However, the more pressing question should be, "What accounts for this racial-based difference?" Psychometric, environmental, and idiosyncratic factors aside, the first probable factor is limited local control over test difficulty levels and related standards of the norm group, and the second is the availability and use of resources as applied to the subject matter tested. Such pitfalls associated with summative norm-referenced testing reinforce the criterion-referenced model currently used in Alberta, which is also being used more frequently in the United States.

Another charge often levied against norm-referenced achievement tests is that they not only exclude certain cultural, racial, and socioeconomic groups, but might also be biased against special needs students. This category is simply a different manifestation of the previous argument against high-stakes, norm-referenced achievement testing. However, instead of being exclusionary based on culture, or socioeconomic factors, the tests are thought to be unfair towards students with special learning needs (as well as language barriers, etc). Ysseldyke and Salvia (2001) isolate five main factors that affect accurate assessments of disabled students, which teachers and administrators should consider when administering, designing, and evaluating their high-stakes achievement tests results:

1. *The ability to understand assessment stimuli*: If students cannot understand the test because of a disability (like sight, hearing, or language) their performances on the tests should be considered a reflection of physical or sensory limitations. Likewise, if a test is translated to accommodate language, validity must be established in the new language.
2. *The ability to respond to test stimuli:* If physical or sensory limitations inhibit accurate responding, the tests are invalid.
3. *Nature of the norm group:* If the test is administered to an individual differently than to the

"norm group" he or she will be referenced against, the comparisons are invalid.
4. *The appropriateness of the levels of items:* Out-of-level tests are inappropriate.
5. *Exposure to the curriculum being tested:* If the students have not been exposed to the curricular content needed to respond, poor performance simply reflects a "lack of opportunity to learn."

Because of such concerns, the American Educational Research Association (AERA, 1999) with several partner organizations published a set of guidelines for school district authorities and policy-makers to consider when implementing high-stakes achievement tests. The AERA guidelines note that it is imperative to ensure that the tests accommodate students with disabilities and do not merely reflect language proficiency unless that is a primary purpose of the test. Canadian schools have as a similar reference, the document "Principles of Fair Assessment Practices for Education in Canada" (Joint Advisory Committee and Working Group, 1993), produced by a number of provincial and territorial ministries and departments of education, working with a joint advisory committee comprised of various stakeholders. It outlines, for both classroom teachers and designers of achievement tests, general principles for fair assessment practices, with a series of general guidelines for each. The principles pertain to developing and choosing methods for assessment, collecting assessment information, judging and scoring student performance, summarizing and interpreting results, and reporting assessment findings. As a result of organizations like these issuing guidelines, Cizek (2001, p. 23) claims, "recent federal legislation enacted to guide the implementation of testing has been a catalyst for increased attention to students with special needs." In short, the act of implementing tests may have brought to the fore an awareness of special needs testing issues that might have otherwise gone neglected, and can thus be viewed as having a positive effect.

Gender is another area where some claim a bias exists within high-stakes achievement testing programs. Easton and Cowley (1999) looked at the differences in British Columbia classroom achievement test grades between boys and girls, and found that girls received better school grades than boys. However, the question that should be asked regarding the difference is whether it represents item bias or a true difference inherent in the subgroups of the population. Gierl, Khaliq, and Boughton (1999, p. 3) define item bias as systemic error in "how a test item measures a construct for the members of a particular group." Therefore, the real issue in such a situation is recognizing differential item functioning, or the likelihood that two different groups have differing probabilities of answering the same question correctly when other factors are controlled for, and interpreting the results bearing that in mind. Further, it is worth noting that a study done by Alberta Learning (2000) on differences between mean scores of males and females on the criterion-referenced provincial achievement tests found that gender explained very little of the variation in scores, and the small amount of variance that gender explained was irrelevant to the measurement of the construct. Thus, the case can be made that there is no gender bias in achievement tests that are properly designed, criterion-referenced, and well fitted to their purpose. Additionally, gender is blind to the centralized markers of the achievement tests in Alberta; thus, there can be no bias based in gender where the expectations of the markers are concerned. Bearing this in mind, the Alberta achievement tests could be a useful control group for further studies of gender effects.

Are teachers being de-skilled by high-stakes achievement tests? At the core of this issue is a belief that teacher's assessment skills and responsibilities are eroded, or atrophy because they are not used when high-stakes achievement testing is implemented. Runte (1998, p. 166) notes, "A case study of provincial testing in Alberta, however, reveals that teachers have retained their assessment skills and their responsibility for evaluation technique, and so cannot be said to have undergone de-skilling." Instead, Runte claims that teachers have lost control over their work processes, as they no longer have the right to evaluate and define their labour practices. In this respect, he claims that the professional status of teachers has been undermined, resulting in an "ideological proletarianization" and a shift from student-centred to curriculum-centred instruction. However, this potential phenomenon can be eliminated if student assessment models that incorporate high-stakes achievement testing are fleshed out with school-awarded marks, as is done with the Alberta diploma exams, whereby a student's final mark is a 50–50 ratio of the achievement test result and the school-awarded mark. This blended model reinforces the importance of analyzing the relationships between the school-awarded and the high-stakes achievement test results in order to fully inform insight into students' learning, and holds implications for more optimal models of relating the grade 3, 6 and 9 provincial achievement tests to classroom-based assessment data.

As a final note, there is one additional political implication that should be considered with the publication of test results. Student results become a form of political currency when the Ministry of Learning announces achievement test scores. In other words, when government announces achievement based on student results with the intention of informing the public's view of the ministry and education system, the issue has become politicized. This in itself is not a problem, as such issues are political, and the public has a right to know what is happening with important institutions/systems that have wide-reaching impacts on everyone's lives. The problem arises when the issue gets reduced to political rhetoric. The debate regarding high-stakes achievement testing can be full of jargon and emotion and could therefore easily degenerate to a level of limited and limiting dialogue. Given this, and the already negative view of high-stakes achievement testing held by some educators and others, it seems imperative that the use of test results be well scrutinized, and the reasons for testing and communication strategies incorporate the delimitations and limitations built into the results base being reported.

In summary, the main effects to consider in the category of political issues are summarized in Table 3.

Table 3: Accountability and Political Implications

(+) Positive Implications	Main Issues: Specific to Bias in Testing and Use of Results	(-) Negative Implications
If tests are criterion-referenced, and properly designed, there is limited bias against SES, gender, and special needs students	Nature of test score interpretation	If tests are norm-referenced, and poorly constructed, bias is more likely
Formative: used to inform curriculum	Administrative and	Summative: used to

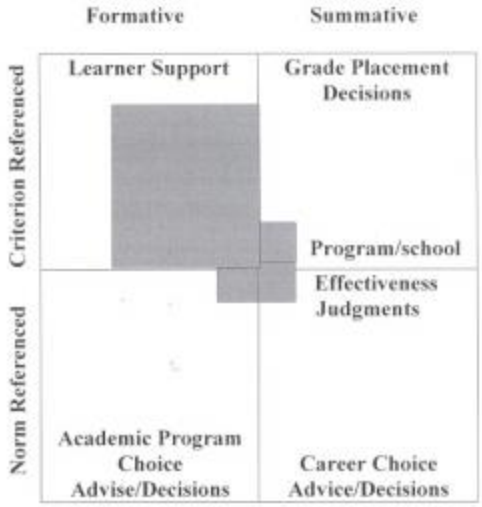
and classroom practices specific to individual performance	classroom use of results	evaluate and label without underlying analysis
Testing and communication strategies include delimitations and limitations in results being reported	Political use of results	Testing issues reduced to political rhetoric

Summary

We should resist the temptation to conclude that the two sides in the high-stakes achievement testing debate have sufficiently opposite views so as to be mutually exclusive or to make a meaningful synthesis impossible. Indeed, the goal of this research review is to find the points where the arguments converge as well as diverge, and to use those insights to help build a system of achievement testing that focuses on the goal of continuous improvement while serving the needs of administrators, teachers, students, and parents alike. In reality, this may not be as difficult as it seems.

In the opening paragraphs of this article, it was claimed that the sides were divided on the basis of their value assumptions. Yet, the factors guiding those values could become the common thread where both groups concur. No one group is involved in the debate with malicious intent in mind: rather, each has involved itself with the goal of creating the best possible education experience for students. Thus, the somewhat adversarial roles often adopted by the groups may not be mutually exclusive constructs. Policy-makers and administrators do not likely wish to implement systems of high-stakes achievement testing to make "life hard" for teachers, nor do teachers likely express concerns for the sole reason of voicing opposition for opposition's sake. In reality, the two are polarized only by their variant beliefs in what is the best way to do the best thing for all involved. In this regard, there is little dispute that the need to continually improve and adapt the learning system is a laudable objective. As long as the common goal is to create an education system that effectively and efficiently serves the needs of those involved in it, reconciling the two sides is a matter of finding the balance. More to the point, it may be a matter of convincing each side that a balanced approach to student assessment exists. This balanced approach to testing generally must give appropriate play to the multiple functions of classroom assessment where testing may be formative or summative and criterion-referenced or norm-referenced relative to varying purposes of testing. Classroom assessment that is based on formative, criterion-reference testing is a powerful teaching tool (Bloom, 1980; Stiggins, 2001), and as such should comprise the majority share of classroom assessment practices. Other forms of testing, such as summative, norm-referenced testing, also have their place in filling out the picture of student achievement and providing appropriate feedback to student, parents, and policy-makers. This holistic but integrated approach to testing is represented in Figure 3, which suggests that the purposes for testing legitimately exist in all quadrants simultaneously, but that the ratio of emphasis and time devoted to varying approaches to classroom assessment should be biased in favour of serving the needs of the student.

Figure 3. A balanced model of classroom assessment



The overarching principal or philosophical concept that relates to all of the purposes and responses to high-stakes achievement testing mentioned in this article is "fairness." Likewise, both sides react so vehemently because fairness and related concepts of what is or is not fair direct so much of the debate in relationship to how testing is used in different contexts. When policy-makers promote high-stakes achievement testing as a means of supporting accountability and making decisions, they are essentially trying to develop a fair process for judging program effectiveness and stimulating improvement. When opponents attack test methodologies, they do so because they believe the tests are not fair measures, and when people argue that tests create social divisions, they claim fair learning opportunities do not exist. It is quite possible that fairness is best ameliorated by achieving consensus on the appropriate balance of testing purposes represented in Figure 3.

AERA's *Standards for Educational and Psychological Testing* (1999) cites four general views of fairness that are most often found in literature pertaining to high-stakes achievement testing. Heubert and Hauser (1999) summarize these by claiming the first two views of fairness are concerned with the absence of bias and equal opportunity for all examinees in the testing process, which speaks directly to the validity and reliability issues raised in the section on methodology. The next view sees fairness in terms of opportunity to learn, and thus informs much of the ideological and political debates, as does the final view of fairness, which is concerned with equality of testing outcomes in the political context.

As Taylor (2001, p. 5) notes, "generally speaking, the provinces with the best-developed assessment cultures have the strongest results in national and international comparisons of student achievement." This is not to suggest the advocates of high-stakes achievement testing have won, but it does give credence to the fact that there is a benefit of critical reflection to systems that employ some method of high-stakes, criterion-referenced achievement testing. Further, this should not be taken as a final approach to the subject, which suggests that high-stakes achievement testing is not immutable. High-stakes achievement testing is merely one component in a system of continuous improvement that strives to further fine-tune, adapt, and

support education. Although high-stakes achievement testing may be a valuable component of such systems, there are numerous emerging technologies that could be incorporated into student assessment practices in order to advance their efficacy and establish meaningful and useful linkages to ongoing classroom-based assessment.

This paper has attempted to assess the positive and negative, intended and unintended impacts of high-stakes achievement testing. Doing so, it is hoped, will make it possible to isolate and manage those aspects that contribute positively to the development of a fair, comprehensive, and integrated model of classroom assessment where high-stakes achievement tests are an integral component of and contribute to the broader classroom-based assessment processes.

***Note*:** The views expressed in this article are those of the authors and not necessarily Alberta Learning's.

Endnotes

1. For an example of this see the National Center on Educational Outcomes website at www.coled.umn.edu/NCEO.
2. Testing in Alberta begins at Grade 3.
3. In the mid-1980s, Ross Perot introduced a number of education reforms in Texas that have since become the Texas Assessment of Academic Skills (TAAS) system, more colloquially referred to as Texas Accountability System. Perot's reforms focus on the use of standardized tests to centralize decision-making for students, teachers, and administrations. In short, the reforms strengthened bureaucratic controls by employing business/management accountability measures.

References

AERA. (2000). *Guidelines for high-stakes tests. Education World* (School Issues Article). Available online: http://www.education-world.com/a_issues/issues110.shtml

AERA. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Alberta Learning. (2003). *Provincial achievement tests: Supporting excellence in student learning in Alberta.* Available online: http://www.learning.gov.ab.ca/k_12/testing/achievement/supporting_excellence/default.asp.

Alberta Learning. (2000). *Performance differences of males and females on Alberta provincial tests*. Edmonton: Author.

Bloom, B. (1980). The new direction in education research: Alterable variables. *Phi Delta Kappan, 61*(6), p.382-385.

Casas, F.R., & Meaghan, D.E. (2001). Renewing the debate over the use of standardized tests in the evaluation of learning and teaching. *Interchange, 32*(2), p. 147-181.

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

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of educational opportunity*. Washington, DC: National Center for Educational Statistics.

Compas Inc. (2001). *The educational experience: Public opinion on education.* Available online: http://www.compas.ca/html/archivesdocument.asp?offset=100&compasID=270

Dounay, J. (2000). High-stakes assessments bring out the critics. State Education Leader, 18(1). Available online: http://www.mccte.msu.edu/news/policy/other/elwinter.pdf

Easton, S., & Cowley, P. (1999). *Boys, girls and grades: Academic gender balance in British Columbia's secondary schools.* Vancouver: Fraser Institute.

Gierl, M., Khaliq, S., & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: Prevalence and policy implications.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Sherbrooke, Quebec.

Haney, W. (1999). *Study of Texas Education Agency statistics on cohorts of Texas high school students, 1978–1999*. Unpublished paper. Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Hess, F.M., & Brigham, F. (2000). The promise and peril of high-stakes testing. *American School Board Journal, 187*(1), p. 26-29.

Heubert, J.P., & Hauser, R.M. (Eds.). (1999). *High stakes: Testing for tracking, promotion and graduation.* Washington, DC: National Academy Press.

Joint Advisory Committee and Working Group. (1993). *The principles of fair assessment practices for education in Canada.* Edmonton: Because We Care Society of Alberta.

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

Lewis, A. (2000). *High-stakes testing: Trends and issues.* Aurora, CO: Mid-Continent Research for Education and Learning, Policy Brief.

McNeil, L. (2000). Creating new inequalities: Contradictions of reform. *Phi Delta Kappan, 81*(10), P. 729-734.

OECD (Organisation for Economic Co-operation and Development). (2001). Measuring up: The performance of Canada's youth in reading, mathematics and science. *Programme for International Student Assessment.* Ottawa: Human Resources Development Canada.

Paris, S., & Olson, G. (1983). *Learning and motivation in the classroom*. Hillsdale, NJ: Erlbaum.

Paris, S., Lawton, T., Turner, J., & Roth J. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher, 20*(5), P. 12-20 & 40.

Popham, J. (2001). Teaching to the test? *Educational Leadership, 58*(6), p.16-20.

Public Agenda. (2002). *Reality Check 2002.* Available online: http://www.edweek.org/ew/newstory.cfm?slug=25realitycheck.h21

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

Rogers, T. (1997) *Examination on the influence of selected factors on performance on Alberta Education Achievement Tests within Edmonton public schools.* Center for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform* (1st ed.). San Francisco: Jossey-Bass.

Runte, R. (1998). Tthe impact of centralized examinations on teacher professionalism.*Canadian Journal of Education, 23*(2), p. 166-181.

Saunders, W.L., & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Schofield, J. (2001). Saving our schools. *Maclean's Magazine, 114*(20), p. 22-28.

Simon, B. (1993). *The Education Reform Act: Causative factors.* In BERA Assessment Policy Task Group (eds.), Policy issues in national assessment. Philadelphia, PA: Multilingual Matters Ltd.

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

Stiggins, R.J. (2001). *Student involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Taylor, A.R. (2001). Policy watch: Student assessment. Education Analyst? *Society for the Advancement of Excellence in Education,4*(3). Available online: http://www.saee.bc.ca/TaylorReport.pdf

Worthen, B.R., & Spandel, V. (1991). Putting the standardized test debate in perspective. *Educational Leadership, 48*(5), p. 65-69.

Ysseldyke, J.E., & Salvia, J. (2001). *Assessment* (8th ed.). Boston & New York: Houghton Mifflin.

Author Notes

Dr. John M. Burger is a Senior Manager in the System Improvement and Reporting Division, System Improvement Group, Alberta Learning. Email: john.burger@gov.ab.ca John is active in implementation of the provincial accountability model in the basic and post-secondary sectors. John also holds an Adjunct Associate Professor appointment at the University of Calgary and teaches distance learning, graduate levels courses in Classroom Assessment.

Mr. Monte Kruger is a Research Officer in the System Improvement and Reporting Division, System Improvement Group, Alberta Learning. Email: monte.krueger@gov.ab.ca Monte completed his Master's Degree in Political Science at the University of Calgary. His primary interests are in the area of policy analysis. His current responsibilities are focusing on assessment policy implementation, social promotion policy and an investigation of the costs of implementing technology in schools.