

Purposes of and Issues
with the
Provincial Testing Programs
in Alberta
W. Todd Rogers
University of Alberta
Donald A. Klinger
Queens University

Mailing address:

Dr. W. Todd Rogers
Centre for Research in Applied Measurement and Evaluation
Faculty of Education
University of Alberta
Edmonton, Alberta T6G 2G5 Canada
E-mail: Todd Rogers <torogers@maildrop.srv.ualberta.ca>

W. Todd Rogers (PhD) is a Professor, Centre for Research in Applied Measurement and Evaluation, University of Alberta in Canada. He is interested in research in the area of measurement and evaluation of students with a focus on large-scale testing. He is particularly interested in the fair assessment of students. He was the chair of the group that developed the *Principles of Fair Student Assessment Practices for Education in Canada*, and he worked on a project to develop the *Student Evaluation Standards* in the United States.

Donald A. Klinger (PhD) is an Associate Professor

We wish to thank the teachers, principals, and superintendents in Alberta who completed the questionnaire. We also wish to thank the Social Sciences and Humanities Research Council of Canada for supporting this research, Alberta Education for facilitating the administration of the teacher questionnaire, and the College of Alberta School Superintendents for facilitating the administration of the superintendent and principal questionnaires. The results and views presented in the paper are attributable to the authors and do not to the agencies that supported and facilitated the research.

The fear of “declining standards” and other indicators that schools are not doing a good job of quality control have been expressed by the public over the past 30 years. These fears continue through today. Politicians responsible for allocating funds for education want to know the impact or results of their investment. Embraced by both federal political parties, every student in the United States is tested at the end of Grades 3 through 10 as a major part of that nation’s effort to use performance standards and assessments as tools for raising student achievement through holding schools accountable. Similarly, we are now witnessing in Canada new and expanded assessment programs in Canada that hold schools accountable in an effort to increase student achievement. Consequently, testing has changed from an instrument for decision-making about students (e.g., unit/chapter test scores grades and grades to measure student progress; selection to tertiary education institutions) to that of a lever to hold schools accountable (Firestone, Mayrowetz, & Fairman, 1998).

As early as 1985, McLean (1985) noted that provinces and school districts in Canada were turning to examinations as a substitute to consultants and inspectors for quality control (McLean, 1985). More recently, large education reforms have explicitly incorporated either the use of new assessments or the expanded use of existing ones. For example, grade 12 Provincial Examinations, which account for 40% of the final Grade 12 student course marks, were introduced in British Columbia in 1984. In 1999 British Columbia added its Foundation Skills Assessment Program for students in Grades 4, 7, and 10 with student results reported to parents and school results used as a school indicator. The Grade 10 FSA was dropped in 2003 and replaced by a new set of exams that count 20% of students final year end marks: English, Mathematics, and Science at Grade 10 and Social Studies at Grade 11. In Alberta, province-wide assessments at Grades 3, 6, 9, and 12 were introduced in 1982. In 1996, the number of assessments conducted at each grade level was increased, progress over time became part of reporting, and school level reporting to parents and the public became mandatory. In Ontario, provincial assessments at Grades 3, 6, and 9 were introduced in 1996, and beginning in 2002, students in high school will need to successfully complete a grade 10 literacy

assessment before they can graduate. There appears to be an unquestioned faith in using assessments in this manner with limited explicit conceptual foundation on which to justify such use.

Typical Concerns

Concerns have been raised, particularly in the United States and Britain, about the soundness and validity of large-scale, high-stakes assessment programs and their use for accountability. To understand these concerns, it is useful to first look at the character of assessments used for accountability. Briefly, these assessments are mandated by an external (to the classroom) agency; are marked by a high-pressure environment; typically are formal in nature to ensure reliability; are widely applicable across schools, classes, and students; are sensitive to aggregated change; and are centrally processed. Results are often of a summative nature and must be in a form useful to policy makers, which usually means reducing complexity to, preferably, a single score for a school, district, or province/state amenable to the ranking of school districts and schools (Rogers, 1991).

Research has consistently found that classroom teaching and assessment practices are influenced by high-stakes large-scale assessments; however, these effects may not be desirable. Darling-Hammond, Aness, and Falk (1994) pointed out that despite the apparent psychometric qualities of such assessments, their impact was such that curriculum was narrowed, instructional time was reduced in favor of test preparation activities, teaching practices were more test like, and cheating increased. Shepard (1991) noted increased test scores were not necessarily indicative of higher achievement. These concerns have not subsided. Several papers debated at the 2000 Annual Meeting of the American Educational Research Association dealt with concerns related to inequities among student sub-populations (Sessions 1.21, 3.53, 23.45), inappropriate test preparation (Session 5.60), adverse influences upon instruction and school improvement (Sessions 11.38,13.40, 15.54, 17.49), inappropriate impact on and evaluation of school personnel (Sessions 13.17, 13.30, 39.02, 46.25, 50.30), incorrect emphasis on testing lower rather than higher order thinking skills (Session 17.38), and problems in setting

standards (Session 27.24, 48.17). In the general session, “Three Blueprints for a Revolution: How to Halt the Harm Caused by High-Stakes Tests” (Session 37.27), Berliner, Popham, and Shepard debated large-scale assessments and their use (AERA Program, 2001).

Dorn (1998) argued that the existence of high stakes large-scale assessments is due to political cultures that have not been examined closely by critics. Rather, studies in which concerns about large-scale assessments are examined typically focus more on issues of psychometrics, equity, technology, and form, rather than the role external assessments play in the larger society. As suggested by Delandshere (2001):

Current assessment practices are debated with regard to the impact they have on defining or narrowing the curriculum, for example, or on the need for different forms of assessment (e.g., Shepard, 1989). Only in rare cases have scholars examined the functions that assessment serves in validating and reproducing certain forms of knowledge or ideas at the expense of others (e.g., Hanson, 1993; Kvale, 1993, Broadfoot, 1996). (p. 115).

While these latter concerns and issues may be debated in the public media, teacher associations, and other interest groups, they have not been examined by the community of measurement specialists responsible for the development of large-scale assessments and the reporting of results.

The fundamental nature of educational assessment is to make judgments about what students know or can do and, increasingly, how schools perform (Delandshere, 2001; *The Principles for Fair Student Assessment Practices for Education in Canada*, 1993). However, what constitutes the basis of these judgments? The beliefs, assumptions, and ideologies of those making the judgments are rarely discussed or made public. Rather, the perspectives taken are assumed to be understood and agreed upon. There appears to be a common belief that assessment is primarily a matter of technique and procedure (for which there have been important developments [e.g. machine scoring;

item response theory; computer-adaptive testing]) to which other, important and telling concerns, are subordinated. Wilbrink (1997), in his historical review of assessment, concluded “It is fascinating to observe that assessment procedures handed down by tradition were in this century uncritically adopted in mass education, possibly leading to major inefficiencies in education and, for too many students, a lack of quality of school life” (p. 44). Delandshere (2001) went on to ask “Is it simply the case that the dominant paradigm and practices remained unquestioned because they are assumed to be agreed upon by the education community? Is it the case that we now have generations of educators, test developers and users who function within an assessment tradition, using its methods, concepts and standards, without really knowing how this tradition originated and evolved?” (p. 119). While the quality of education and schooling has received much criticism, to the possible detriment of students, the means by which we judge that quality has received little.

High stakes assessment programs have strong consequences and research is needed in Canada as well as elsewhere to evaluate the degree to which these assessments meet expectations (Baker, O’Neil, & Linn, 1994). A recent special issue of *Educational Measurement: Issues and Practice* on “Test Scores and State Accountability” (2 xxx, Volume 24, Number 4) looked state testing and accountability that arose as a result of the implementation of No Child Left Behind requirements. After discussing the four articles included in this special issue, Hess (2005) concluded “while scholarly scrutiny will not necessarily settle debates, it can help yield more constructive and informed [policy and accountability] decisions. In particular, this research can clarify actual consequences of policy decisions; highlight and refine approaches that may be more reliable, stable, and effective than those in use; flag the unanticipated or overlooked effects of design decisions, and ensure that both policymakers and the public are aware of the costs and benefits of accountability” (p. 53).

In 1994, Cheliminsky and York (1994), on behalf of the American General Accounting Office in the United States, examined large-scale assessment programs in Canada. They concluded that British Columbia and Alberta, in particular, had exemplary

programs that should serve as models for future assessment programs in the United States. However, since that time, the purposes and uses of these programs have expanded and changed, particularly in the use of the results for accountability purposes. There has been little research examining these expanded and new purposes. Consequently the intent of the present study was to examine the views of teachers, principals, and school district superintendents in Alberta regarding the explicit and implicit purposes for which these assessments are developed and implemented and the issues and concerns that have been raised about these assessments.

Method

The Alberta Education, the Department of Education, has two testing programs: the Provincial Achievement Testing (PAT) Program and the Diploma Examination (DE) program. PATs are administered at Grade 3 (Language Arts, Mathematics) and Grades 6 and 9 (Language Arts, Mathematics, Science, Social Studies). Teachers are encouraged, but not required, to include the marks their students achieve on these tests as part of the final mark for the school. Principals and Superintendents are required to report the results to the public. Alberta Education also sends individual student results directly to their parents. DEs are school leaving examinations administered in all senior English, mathematics, science, and social studies courses. The results from these examinations are combined with the teacher awarded marks so that each accounts for 50% of the total final mark. The items included in the PATs and DEs are based on the knowledge and skills that are set out in the Programs of Studies teachers are expected to use and follow.

Identification of Explicit and Implicit Purposes of the Assessments

Representatives of the Learning Assessment Branch, Alberta Education and the Alberta Teachers' Association were interviewed to elicit the explicit and implicit purposes of the Provincial Achievement Tests and the Diploma Examinations and to identify the issues associated with each of these testing programs. Two senior officials with Alberta Education were interviewed together and a senior staff member and three

staff members with the Alberta Teachers' Association were interviewed as a group. These two arrangements were set by each organization. The following two major questions were asked, one related to the purposes and uses of the two programs and the second to issues with each program, were used to focus each interview:

1. What are the purposes and uses of the
 - a. Provincial Achievement Tests (PATs) administered at Grade 3 (Language Arts and Mathematics) and Grades 6 and 9 (Language Arts, Mathematics, Science, and Social Studies)?
 - b. Diploma Examinations for each Grade 12 course (e.g., English 30 and 33, Pure Mathematics 30 and Applied Mathematics 30)

2. What issues do you see today with respect to the
 - a. PATs
 - b. Diploma Examinations?

Each assessment program was considered separately for each question. The participants were reminded to consider both explicit and implicit purposes. Probe questions included seeking clarification or an example. Each group interview required approximately 90 minutes to complete.

Interview results. Given, the purposes identified from these interviews were used to develop the questionnaires to be used with teachers, principals, and superintendents, the results of the interviews are discussed here. Table 1 contains a listing of the purposes/uses and issues about the provincial achievement testing and diploma examination programs. Interestingly, the purposes/uses and issues identified in both interview sessions were essentially the same. This finding is attributable to the age of the program (started in 1982) and the time since the last revision. What was different is the valence attached to some of the purposes and issues. For example, the ATA felt that a provincial test at the third grade level was inappropriate given the age of the students

while ABED felt that the third grade students did not experience trouble responding to the test items. Both groups indicated that there needs to be a balance between classroom assessments and external standardized tests and that some standardized measure is necessary to check that the educational system is working as intended. At the very end of the meeting with the government officials, the senior government official provided a copy of a document sent out to the superintendents entitled “Genesis of an Accountability Pillar.” Of relevance here is the role of the Diploma Examinations and the PATS: they are considered to be “critical measures for the accountability pillar.”

 Insert Table 1 about here

Development of Questionnaires

Three questionnaires – teacher, principal, and superintendent – were constructed using the information provided in Table 1, from the questionnaires used in an evaluation of the provincial examination program in 1989 (Anderson, Muir, Bateson, Blackmore, & Rogers, 1990), from a review of the test bulletins distributed by Alberta Education to schools in the Fall and the web-site maintained by Alberta Education (www.education.gov.ab.ca/). Each questionnaire contained four sections. The first section contained questions related to teacher assignment (teachers), school (principals), and school districts (superintendents). The next three sections were identical. The second section contained questions about the appropriateness of purposes and uses of the provincial assessments. Separate lists of the purposes and uses were provided for the provincial achievement tests and the diploma examinations. Teachers completed one or both of these subsections depending on what grades that they taught; principals completed one or both sections depending on what grades were in their schools; and superintendents completed both sections. In the third section, all participants rated the seriousness of issues raised in the education community about the provincial achievement tests and the diploma examinations and indicated at what grade levels provincial assessments should be administered. A five-point Likert scale was used to assess the degree of the appropriateness of the purposes/uses of the PATs and DEs (1 = not

appropriate, ... , 5 = very appropriate) and the seriousness of the issues with the PATs and DEs (1 = not an issue, ... , 5 = very serious issue). To whom results should be reported and the grade levels at which province-wide testing should be conducted were measured using a two-point yes-no format. The fourth section contained bio-demographic questions (gender, educational qualifications) and questions about how confident the participants were about you about their knowledge of the two provincial assessment programs and the national and international assessments administered in Canada (Note 1).

Administration of Questionnaires

The teacher questionnaires were administered to the teachers involved in marking the provincial achievement tests and diploma examinations during the summer, 2005. It was felt that these teachers would have good knowledge about the purposes and uses of the achievement tests and diploma examinations and the attendant issues. In all subjects except of Grade 12 English, time was allocated to complete the questionnaires during the marking session. In the case of the Grade 12 English, questionnaires were distributed to the teachers who were asked to complete it outside of marking time and then to return in to a box placed at the front of the marking room. Approximately 20 minutes was required to complete the questionnaires.

The superintendent and principal questionnaires were administered using e-mail. The Executive of the College of Alberta School Superintendents endorsed the study and sent a copy of the superintendent and principal questionnaires to the superintendents by e-mail and asked the superintendents to forward the principal questionnaire to the principals in the school district. Completed questionnaires could be returned using e-mail or the regular surface mail. This procedure was altered in one district. District officials felt it better to send the questionnaires using surface mail. The data collection for superintendents and principals started in January 2006 and was terminated in June 2006 (Note 2).

Analysis

Data entry. A bonded data entry company entered the rated and coded responses into a computer file with 100% verification. The researchers reviewed the comments that were made in response to the open ended items. Since the comments did not add to results of the analysis of the rated and coded data, they were not coded. Instead, representative comments were selected to illustrate the quantitative results.

The individual purposes and issues were grouped according to their common function (e.g., assessment of learning, improvement of curriculum and instruction). The responses of the teachers in the Grade 3, 6, and 9 PAT samples were first compared to determine in these three samples could be collapsed into one. A one-way ANOVA, employing the Browne-Forsythe (1974) test statistic given the lack of homogeneity of variance for some of the items and unequal sample sizes, was then conducted for each item. Simultaneous pair-wise multiple-comparisons tests employing Tamhane's (1979) procedure for samples of unequal size were completed for each item for which a significant difference was indicated. Given the exploratory nature of this study and the belief that the consequences of Type II error would be more costly than the consequences of a Type I error, all analyses were completed at the 0.05 level of significance. The analyses were completed using Version 14.0, Statistical Package for the Social Sciences (SPSS, 2006). The findings revealed that while there were statistical differences, the effect sizes were all small (≤ 0.49 ; Cohen, 1988). Consequently, the Grade 3, 6 and 9 samples were combined.

Scale means, standard deviations, and measures of internal consistency were then computed for each group of items and the samples of teachers, principals, and superintendents who responded for the PATs and who responded for the DEs. Given homogeneity of variance, one-way ANOVA was then conducted for each scale. Simultaneous pair-wise multiple-comparisons tests employing Tukey's (1979) procedure were completed for each item for which a significant difference was indicated. The harmonic mean of the samples sizes was used in these analyses. As well, effect sizes (ES) were determined for those comparisons found to be significant. Cohen's (1988)

suggestions for small ($0.20 \leq ES < .50$), medium ($0.50 \leq ES < 0.80$), and large ($ES \geq 0.80$) were adopted. Cases for which there was lack of transitivity were not claimed (Note 3). Given the exploratory nature of this study and the belief that the consequences of Type II error would be more costly than the consequences of a Type I error, all analyses were completed at the 0.05 level of significance. The analyses were completed using Version 14.0, Statistical Package for the Social Sciences (SPSS, 2006).

Results

The teachers who marked the provincial achievement tests and the diploma examinations were classified by the grade level at which they marked: 122 teachers marked the Grade 3 Language Arts Writing, 150 teachers marked the Grade 6 Language Arts Writing, 246 teachers marked Grade 9 Language Arts Writing, and 475 teachers marked the Grade 12 constructed responses in English Language Arts, French Language Arts, Social Studies, Mathematics, Biology, Chemistry, Physics, and Science. Likewise the principals were classified, but this time according to what tests/examinations were administered in their schools: 129 PAT principals in schools with at least one of Grades 3, 6 or 9, 35 DE principals of schools with Grade 12. The total number of school district superintendents who responded was 23 (Note 3).

Description of Teachers, Principals, and Superintendents

and their Schools and School Districts

Teachers. As shown in Table 2, the percentage of male teachers increased with increasing grade; while 21.2% of the Grade 3 teacher markers were male, 57.8% of the Grade 12 teachers were male. While the proportion of teachers with a B. Ed. tended to decrease with increasing grade, the proportion of teachers with an undergraduate degree plus a B. Ed. or with a M.A./M.Ed. tended to increase with increasing grade.

Insert Table 2 about here

The teachers were asked to describe the school in which they taught in terms of the socio-economic level of the community the school served, the academic standing of the school relative to other schools, and the location of the school. The results are summarized in Table 3. While greater proportions of Grades 9 (35.6%) and 12 (41.7%) teachers than Grades 3 (19.8%) and 6 (24.5%) teachers indicated that the school in which taught served a community whose socio-economic level was above/far above the average total income (before taxes) in Canada (\approx \$64,000), greater proportions of Grades 3 (43.0%) and 6 (37.4%) teachers than Grades 9 (22.6%) and 12 (18.5%) teachers indicated that they taught in school that served a community whose socio-economic level was below/far below the Canadian average. The teachers provided a similar profile when they described the academic performance of their schools relative to other schools. A greater proportion of Grades 9 (44.0%) and 12 (42.3%) teachers than Grades 3 (35.6%) and 6 (28.0%) indicated that the academic achievement of their schools was above/far above average. In contrast, a greater proportion of Grades 3 (23.9%) and 6 (22.7%) teachers than Grades 9 (12.3%) and 12 (12.2%) teachers felt the relative achievement of their schools was below/far below other schools.

Almost equal proportions of Grade 6 (38.9% and 41.6%) and 12 teachers (47.1% and 47.0%) taught in schools located in the inner city/urban area and in the semi-rural/rural areas. In contrast, while a greater proportion of Grade 3 teachers taught in inner city/urban areas than in the semi-rural/rural areas (52.0% vs. 34.2%), a greater proportion of Grade 9 teachers taught in the semi-rural/rural areas than in the inner city/urban areas (58.9% vs. 29.7%).

Insert Table 3 about here

Principals. The percentage of DE male teachers was less than the percentage of DE teachers (49.2% vs. 76.2%; see Table 2). The distributions of attained educational level were similar between the two groups of principals, with the most common level being M.A. Med. (66.9% and 61.9%).

The principals were also asked to describe their schools in terms of the socio-economic level of the community the school served, the academic standing of the school relative to other schools, and the location of the school. (see Table 3). Smaller proportions of both the PAT and DE principals indicated that their schools served a community whose socio-economic level was above/far above the average total income (before taxes) in Canada than PAT and DE principals who indicated their schools served a community whose socio-economic level was below/far below the Canadian average (18.3% and 17.7% vs. 40.5% and 35.3%). In contrast, the proportions of PAT and DE principals who reported that the academic achievement of their schools was above/far above average (40.8% and 41.8%) were greater than the proportions of both groups of principals who reported that the achievement of their schools was below/far below other schools (19.4% and 20.6%). Lastly, while more equal, the proportions PAT and DE principals with schools located in the inner city/urban areas were greater the proportions of PAT and DE principals with schools in the semi-rural/rural areas (52.4% and 52.9% vs. 38.1% and 41.1%).

Superintendents. The percentage of male superintendents was 87.0% (see Table 2). less than the percentage of DE teachers (49.2% vs. 76.2%; see Table 2). The majority had a M.A./M.Ed. degree (87.0%), while the remaining 13.0% had a Ph.D/Ed.D.

The superintendents were asked to describe their school districts the academic standing of the district relative to other districts (see Table 3) and the location of the school district. As shown in Table 3, the superintendents indicated that, in comparison to other districts, their district's academic achievement was either at the average (39.1%) level or above average (60.9%) level. A different scale was used for the superintendents given school districts span the locations of schools. The locations of the principals were

30.0% in a large city, 5.0% in a specialized municipality, 45.0% in a municipal district, and 10.0% in each of a town and a village.

In summary, the teachers and principals worked in schools that spanned the range of family income, academic achievement, and locations. In contrast, while the districts represented by the superintendents spanned the range of locations, the superintendents indicated that their districts were at least at the average academic achievement level.

Appropriateness of Purposes/Uses of Provincial Achievement Tests

The purposes/uses of the PATs were grouped into eight functional categories defined by the potential uses of the results (see Table 4). Before proceeding with the presentation of these results, it is noted that the internal consistency values for each category were, with one exception, high, suggesting that the items were consistently measuring the same construct. Further, there are two general findings. First, when a significant difference is interpretable, both the teachers and principals feel less strongly than the superintendents that the purpose/use is appropriate. Second, with the exception of the use of PAT results to improve curriculum and instruction, the Grades 3, 6, and 9 teachers, the principals, and the superintendents did not endorse the purposes/uses of the PATs.

Assessment for learning. Despite the fact that the results are provided to schools prior to the beginning of the next school year, and that teachers spend up to half a day discussing the last year's results before the current school year begins, the teachers, principals, and superintendents indicated that the PATs could not be used to motivate students to work, reduce achievement differences among students, or improve student achievement. While there was a significant difference among the sample means, there was a lack of transitivity. The sample means – 7.15, 6.09, and 8.00 – were all less than the mid-value of score range, 9.00.

Assessment of learning. The teachers, principals, and superintendents felt that the PATs were not appropriate as measures of learning. The sample means – 15.25, 14.20, and 15.98 – were all less than the mid-value score for this category, 18.00, and did not

differ significantly. The teachers, principals, and superintendents indicated that the PATs could not be used to ensure high academic standards, evaluate the quality of students, identify exemplary students, rank students, as part of students' final grades, and how well the students have learned the intended curriculum.

Improvement of curriculum and instruction. As mentioned above, the teachers, principals, and, especially ($p < 0.05$), the superintendents, were more positive about the use of the PAT results to improve curriculum and instruction. The means for this category – 16.01, 14.20, and 18.01 – all exceeded the mid-value, 15.00. The corresponding effect sizes, 0.49 for the teachers and 0.76 for the principals versus the superintendents, were, respectively, small and moderate. Many, but not, all of the schools in the province devote up to half a day discussing the previous year's PAT results before the current school year begins. These discussions generally are related to the curriculum, and how well the curriculum was covered the previous year. This observation is reflected in the variation observed among the appropriateness ratings of the purposes/uses in this category within the three groups. While the respondents in each group were positive about the use of the PATs to support curriculum implementation, focus teaching and instruction on the provincial curriculum, and as a common measure so that teachers can link their own assessments to provincial standards, they were more neutral about the use of the PATs to improve and enhance teaching and instruction and increase teachers' assessment knowledge and skill.

Provide information data for data-based decision-making. The superintendents indicated that the PATs provided results that could be used for decision-making at the student and class level, and more so at the school, school district, and provincial levels. The teachers and principals felt otherwise ($p < 0.05$). The sample means for the teachers and principals – 12.85 and 13.51 – were less than the mid-value of 15.00, while the mean for the superintendents, 16.78, exceeded it. The corresponding effect sizes, 0.84 for teachers and 0.70 for principals, were, respectively, large and moderate.

Provide information for personnel and program evaluation. The use of PAT results for the evaluation of the quality of teachers, schools, school districts, and

provincial education programs and for the identification of exemplary teachers, schools, programs, and school districts was not supported by the teachers, principals, and superintendents. Related to accountability, the sample means – 16.00, 15.22, and 19.17 – did not differ significantly, and all were less than the mid-value, 24.00.

Identify student and school need. The teachers, principals, and superintendents indicated that the use of the PATs to identify students and schools in need was not appropriate. While there was a significant difference among the sample means, there was a lack of transitivity. The sample means – 4.91, 4.20, and 5.52 – were all less than the mid-value, 6.00.

Assessment of progress of cohorts. Cohort analyses are conducted in which the performance at Grade 6 is predicted from performance at Grade 3, and performance at Grade 9 is predicted from performance at Grade 6. The teachers, principals, and superintendents were unsure about the use of the PATS for this purpose. Further, the sample means – 7.60, 7.52, and 9.04 – differed significantly; the means of the teachers and principals were below the mid-value, 9.00, while the superintendents' mean was at the mid-value. The corresponding effect sizes, 0.45 and 0.48, respectively were both small.

Provide information for the parents and the general public. Parents are provided with a report of their child(ren)'s results and school level results are provided in school newsletters and to the general public through publication in local newspapers, and often in the context of school rankings. The teachers and principals, more so than the superintendents ($p < 0.05$), felt this reporting was not appropriate. The sample means for teachers and principals, 7.06 and 7.37, were less than the mid-value (9.00), while the mean for superintendents was just above, 9.17. The corresponding effect sizes were 0.70 and 0.59.

The teachers, principals, and superintendents were asked to identify the three most important purposes/uses of the PATs. Not all teachers, principals, and superintendents ranked the PAT purposes/uses. Further, of those who did, not all identified the top three but only the either first and second or just the first. To summarize

these data, the mean rank for each purpose/use, \bar{R}_G , was computed by taking the sum of the three ratings of importance weighted, respectively, by 3, 2, and 1 and dividing by the number of respondents in each group who provided at least the first ranked purpose. It was assumed that the respondents who did not provide a second and third ranking felt that the other purposes were not important enough to be ranked. The higher the mean rank, the higher the ranking of importance. The most important purpose/use identified by the all three groups was *focus teaching and instruction on the provincial curriculum*. The second most important purpose for the teachers was *provide common measures that teachers can link their own assessments to provincial standards*, for principals *determine how well students are learning the intended curriculum*, and for superintendents *support curriculum implementation*. Lastly, the third most important purpose/use identified by the teachers was *support curriculum instruction*, for principals *provide common measures that teachers can link their own assessments to provincial standards*, and for superintendents *support curriculum instruction*. Each of these purposes was included in the purpose/use category *Improvement of Curriculum and Instruction*, which was rated as the most appropriate of all the categories.

Appropriateness of Purposes/Uses of Diploma Examinations

The purposes/uses of the DEs were grouped into six functional categories defined by the potential uses of the results (see Table 5). As with the PATs, the internal consistency values for each category were, with one exception, high, suggesting that the items were consistently measuring the same construct. When a significant difference was found, principals felt more strongly than the teachers and superintendents that the purpose/use was inappropriate. Second, the Grade 12 teachers and principals, and the superintendents tended to provide more positive ratings than observed for the PATs.

Provide information for post-secondary education. The results of the DEs are combined with the teacher awarded marks so that each counts half of the final course marks for each student. The sample means – 6.64, 6.64, and 6.26 – were slightly greater than the mid-value, 6, and did not differ significantly. The teachers, principals, and

superintendents were essentially neutral about the appropriateness of using the DEs for post-secondary and placement decisions. This result is likely attributable to the use of a blended mark, and the tentative acceptance of Grade 12 students into a post-secondary program prior to writing the DEs. The DEs are, in effect, used to confirm the tentative decision made.

Assessment of learning. The teachers and superintendents, more so than the principals, felt that the DEs were appropriate as measures of learning ($p < 0.05$). While the sample means for the teachers and superintendents – 23.56 and 20.89 – exceeded the mid-value score for this category, 21.00, the principals' mean essentially equaled the mid-value. However, the corresponding effect sizes, 0.45 for the teachers and 0.42 for the superintendents, were small. The teachers, principals, and superintendents indicated that the DEs could be used to ensure high academic standards, certify student competence, help determine final grades, and determine how well the students have learned the intended curriculum. At the same time, they were neutral about the use of the DEs to evaluate the quality of the students, identify exemplary students, and motivate the students to work harder.

Improvement of curriculum and instruction. As with the assessment of learning, the teachers and superintendents were more positive than the principals about the use of the DE results to improve curriculum and instruction ($p < 0.05$). The teachers' and superintendents' means, 17.21 and 17.39, exceeded the mid-value, 15.00, while the principals' mean, 14.74, was slightly less. The corresponding effect sizes, 0.59 for the teachers and 0.63 for the superintendents, were moderate. As previously mentioned, teachers spend up to half a day discussing the last year's DE results before the current school year begins. These discussions generally are related to the curriculum, and how well the curriculum was covered the previous year. This observation is reflected in the variation observed among the appropriateness ratings of the purposes/uses in this category within the three groups. While the teachers and superintendents more so than the principals were positive about the use of the DEs to improve and enhance instruction, as a common measure so that teachers can link their own assessments to provincial standards,

and to ensure fairness in grading across the province, the teachers and superintendents were neutral and the principals less so about the use of the DEs as a balance to the school awarded mark and to increase teachers' assessment knowledge and skill.

Provide information data for data-based decision-making. The teacher, principals, and superintendents were uncertain about the use of DE results for data-based decision-making at the student, class, school, school district, and provincial levels. The sample means– 14.37, 14.77, and 15.74 – bracketed the mid-value of 15.00 and did not differ significantly.

Provide information for personnel and program evaluation. Like the PATs, the use of DE results for the evaluation of the quality of teachers, schools, school districts, and provincial education programs and for the identification of exemplary teachers, schools, programs, and school districts was not supported by the teachers, principals, and superintendents. Related to accountability, the sample means – 18.52, 17.48, and 20.59 – did not differ significantly and all were less than the mid-value, 24.00.

Identify student and school need. The teachers, principals, and superintendents indicated that the use of the PATs to identify students and schools in need was not appropriate. The sample means – 4.66, 3.93, and 4.90 – did not differ significantly and all were less than the mid-value, 6.00.

The Grade 12 teachers and principals, and the superintendents were asked to identify the three most important purposes/uses of the DEs. The procedure used to summarize the PAT importance data (see page 14) was used to summarize the DE data. The three most important purposes/uses of the DEs identified by the teachers were, in order, *ensure the fairness of grading across the province*, *provide common measures so that teachers can link their own assessments to provincial standards*, and *ensure high academic standards*. For the principals, *provide common measures so that teachers can link their own assessments to provincial standards* and *select students for post-secondary education* were tied as most important, followed by *ensure high academic standards* and *ensure fairness in grading across the province*. Lastly, the three most important purposes identified by the superintendents were, in order, *determine how well the students are*

learning the intended curriculum, ensure fairness in grading across the province, and ensure high academic standards and select students for post-secondary education (tied). Taken together, the most important purposes of the DEs were related to ensuring that students are treated fairly for post-secondary decision-making.

Reported Uses of Provincial Achievement Tests and Diploma Examinations

The teachers, principals, and superintendents were asked if they actually used the results and, if so, what they used the results for. Approximately two-thirds (66.4%) of the Grade 3 teachers, slightly more than half (55.3%) of the Grade 6 teachers, approximately 8 out of 10 Grade 9 (79.3%) and Grade 12 (81.9%) teachers, nearly all of the principals (96.0%), and all of the superintendents reported that they did actually use the results of the PATs and/or DEs.

They used the results mostly for improving curriculum and instruction. Approximately three out of five (60.9%) PAT teachers, four of five (78.1%) DE teachers, and all the principals and superintendents who indicated that they used the results identified elements related to the improvement of curriculum and instruction. For example, 41.3% of the PAT teachers and 67.4% of the DE teachers said that they used the results to improve their teaching/instruction and their assessment practices. Another 14.0% of both teacher groups indicated that they compared their school/class results with the provincial results to see how they were doing relative to other teachers. One Grade 12 teacher commented:

I use the results to assess where class where class strengths and weaknesses were to that I can improve my teaching methods and my tests.

The same comment was echoed by a Grade 3 teacher:

I look at areas of strength and weakness, and identify lessons to work on in areas of need.

A quarter of the PAT teachers reported that they used the PATs to determine the final report card marks for the subjects tested. Alberta Education encourages this practice, and when the PAT results are combined with the teacher mark, it is usually at Grade 9, and the most common weightings identified by the teachers were 10% and 25%.

Equal or nearly equal percentages of PAT and DE principals reported they used the results to identify program needs within their school, improve instructional and assessment practices, and develop their Annual School Improvement Plan (28.8%, 28.8%, and 18.9%; 25.0%, 21.9%, and 21.9%). One elementary school principal described the process she used:

As a staff, we analyze the results yearly. Teachers from all grades participate in the analysis. Discussion focuses on how each staff member can contribute to improving how scores. Instructional goals are set at the individual class and school level. Grade 3 and 6 classroom teachers target curricular areas to cover more thoroughly. Consideration is given to the student academic potential of the grade level and the past five year of assessments are analyzed for trends.

Another shared the questions she uses with her staff:

How can we enhance teaching practices?

Were there common questions across the grade that did not go well? Why?

Did the results indicate that we taught the curriculum?

How do the results compare to classroom work and teachers' assessment results?

Did anything change between Grade 3 and Grade 6? What do the cohort results show us?

In contrast, a third elementary principal stated: "I use the results as little as I can professionally get away with." Nearly one in 10 PAT principals used the results to compare their school results to the provincial results (9.9%) and to identify topics for

staff professional development days (8.1%); the corresponding percentages for DE principals were 21.9% and 12.5%.

Over 40% of the superintendents indicated that they used the PAT and DE results to identify program needs (43.5%) and improve teaching and assessment practices (47.8%). Three in 10 used the results to prepare the Annual Education Plan for their school district. Like the high school principals, 21.7% of the superintendents compared the school results to the provincial results and 13.0% used the results to identify professional development topics.

Interestingly, while less than 3% of the teachers reported they used the PATs or DEs to identify students in need, 15.3% of the PAT principals, 9.4% of the DE principals, and 8.7% of the superintendents reported that they used the results for this purpose. Further, 13.0% of the superintendents reported they used the results to identify teachers in need and 17.4% indicated they used the results to identify schools in need. No teachers or principals reported they used the results for these purposes.

Seriousness of Issues with the Provincial Achievement Tests

The issues raised about the PATs and the DEs were classified into five categories. The categories are listed in Table 6 together with their psychometric characteristics. For the first four categories, two sets of teacher results, one set for the PAT teachers and one set for the DE teachers are presented along with the results for principals, undifferentiated according to testing program, and the superintendents. For the fifth category, the results are separated by grade level because the teachers at each grade level tended not to comment about the impact of the curriculum at grades other than their own. The principals are separated into two groups, one of the PATs and the other for the DEs. Before proceeding with the presentation of these results, it is noted that, with five exceptions, the internal consistency values were high, suggesting that the items were consistently measuring the same construct. Again, application of the Spearman-Brown formula to account for the smaller numbers of items in the third fourth categories revealed that the internal consistency values exceed 0.90. Given the issues raised were the same, the results are presented together for both programs.

Characteristics of the PATs and DEs. The teachers, principals, and superintendents were not sure about the seriousness of the issues raised about the characteristics of different components of the PATs. The sample means of the teachers, principals, and superintendents – 21.19, 22.19, and 20.50 – bracketed the mid-value of 21.00 and did not differ significantly. While the sample means for the DEs – 18.73, 22.19, and 20.50 – also bracketed the mid value, they were significantly different. However, there was a lack of transitivity. Within the set of seven issues, concern was evident regarding the inconsistency of performance standards across the years and the limited use of the results due to the security of the test; the respondents were not sure about the reliance on multiple-choice items, and the lack of sufficiency of the results; and felt that that the content and skills assessed each year and the marking of constructed response items were both consistent.

Teachers', principals', and superintendents' knowledge and use of the PAT results. Three pairs of items, related to teachers, principals, and district administrators lack of knowledge about how to interpret and use item and tests results, made up the second category of issues. While the sample means of the three samples – 16.29, 12.14, and 12.39 (PATs) and 16.68, 12.14, and 12.39 (DEs) – were less than the mid-value, 18.00, they did differ significantly ($p < 0.05$). The principals and superintendents felt more strongly that the interpretation and use of results was less of a problem than did both groups of teachers. The corresponding effect sizes, 0.63 and 0.59 for the PATs and 0.74 and 0.70 for the DEs, were of moderate size. Thus, although over all there is no real concern with this issue, the teachers are somewhat more tentative.

Parent, general public, and public press understanding. In contrast to the differences observed for the teachers', principals', and superintendents' knowledge about to how to interpret and use results, there was uniformity among both groups of teachers, the principals, and the superintendents that parents do not know how to interpret the results provided to them, the public press ignores the limitations of publishing school ranks based on the PAT and DE results, and that the general public does not know how to

interpret what is published in the papers. The sample means – 13.19 and 12.81, 12.72, and 12.00 – exceeded the mid-value, 9.00, and did not differ significantly.

Accountability. Both groups of teachers and the principals, more so than the superintendents, strongly felt that the use of the PATs to evaluate the effectiveness of teachers, to publicly rank schools, and as an accountability tool was a serious issue. The sample means – 13.02, 12.63, and 10.85 – exceeded 9.00, the mid-value and were significantly different ($p < 0.05$). The effect size, 0.87, for the PAT teachers compared to the superintendents was large: the remaining effect for the PATs, 0.71. for and the two effect sizes for the DEs, 0.59 and 0.69, were moderate.

Impact on the classroom. Likewise, the teachers and principals at the four grade levels, even more so than the superintendents, indicated the PATs and DEs narrowed the curriculum, reduced instructional time, caused teachers to teach to the test, and limited their own classroom assessments to the item formats used on the PATs. They were unsure about including PAT and DE results as part of the final year-end mark. For example, the sample means at Grade 3 for the teachers and principals – 23.19 and 24.03 – exceeded the mid-value, 21.00; in contrast the mean for the superintendents, 16.64, was smaller. The corresponding effect sizes, 1.12 and 1.26, were large. This pattern of results holds for the other three grades (see Table 6).

Reporting Levels

The results of the Provincial Achievement Tests are reported at different levels in the school system and to the general public. As pointed out above, this is also an issue of concern. To address this, the teachers, principals, and superintendents were asked to indicate to whom they thought student, class, school, district, and provincial results should be reported. The results are summarized in Table 7. Preliminary analyses revealed that the Grade 3 and Grade 6 teachers provided similar responses as did the PAT and DE principals. Therefore these samples have been combined to yield a Grades 3 and 6 teacher sample and a Grade 3 and 6 principal sample. To test the significance of the differences among the proportions shown in each results-report to cell, Marascuilo's

(1996) χ^2 -square test for $k = 5$ independent binomial samples was used to analyze the dichotomous (yes-no) responses. As shown, there are several cells with significant differences, but in each case there is a lack of transitivity. Thus the results are described descriptively.

The cells for which at least half of the teachers, principals, and superintendents indicated that indicated that a report (student, class, school, district, provincial) should be provided to a particular “audience” (students, parents, teachers, principals, public) are set out using dotted lines. Looking first at the student results, the majority of teachers, principals, and superintendents indicated that the student results should be reported to the students, parents, teachers, and principals, but not to the general public. The next level of reporting, class, is somewhat more restricted, with the majority of teachers, principals, and superintendents indicating that teachers and principals should receive class results while students, parents, and the public should not. Like class level results, the majority of teachers, principals, and superintendents stated that the school level results should be reported to teachers and principals. However, approximately six out of 10 (59.0%) principals suggested that the school results should be provided to parents. Further, approximately half (62.2%) of the superintendents indicated that school results should be provided to students and the general public, and seven out of 10 (69.6%) superintendents indicated that students should receive school results. With two exceptions, the pattern for reporting district level results was like the pattern for school level reporting. Less than half of the superintendents suggested that students be provided with the district level results and approximately six of 10 (62.6%) principals indicated that the public should be provided with these results. Lastly, as for school and district level reporting, over half of the teachers, principals, and superintendents indicated that the provincial report should be provided to the teachers and principals. Further, a) slightly less than six out of 10 Grade 9 teachers, Grade 12 teachers, and principals, and approximately 8 out of 10 superintendents suggested that the provincial report should be provided to parents, and, even more, b) half for the Grades 3 and 6 teachers, seven out of 10 (69.1%) Grade 9 teachers, eight out of 10 Grade 12 teachers, nearly eight out of 10 principals (77.0%), and

nine out of 10 superintendents (91.3%) indicated that the provincial report should be provided to the general public. There is no apparent reason why the difference noted between the call for provincial reports differed between parents and the public. In contrast, it is clear that school level reporting to the public that is now in place is clearly not supported by teachers and principals, and by only half of the superintendents (Note 4).

Grades at which Testing should be Conducted

Another contentious issue in Alberta concerns the grades at which province-wide testing should be conducted. The teachers, principals, and superintendents were asked to indicate if the testing done at each of Grade 3, 6, 9 and 12 should be retained (Note 5). They were then asked why they felt the way they did. The percentages of teachers, principals, and superintendents who said yes at each the present grades at which testing is presently conducted are reported in Table 8. The teacher results are for the teachers at each grade level (e.g., 32.6% of the Grade 3 teachers indicated the PATs for Grade 3 should be retained). Likewise, the principal results are for the principals who have the grade in their school (e.g., 86.1% of the principals with Grade 12 in their school reported that the Grade 12 DEs should be retained). The percentages of yes (should be retained) for each grade are based on the number of respondents who answered yes or no for that grade. To test the relationship between respondent group and grade level, a $3 \times 4 \chi^2$ analysis was performed. The result was not significant at the 0.05 level of significance.

As shown in Table 8, the percentages of teachers, principals, and superintendents who indicated that provincial testing should be retained increased with increasing grade level. For example, approximately a third of the Grade 3 teachers indicated that province-wide testing should be retained at Grade 3, approximately half of the Grade 6 teachers indicated PATs should be retained at Grade 6, approximately seven in 10 Grade 9 teachers indicated PATs should be retained at Grade 9, and nine in 10 Grade 12 teachers said that the DEs should be retained at Grade 12. Taken together, the results provided in Table 8 indicate that support for retention of the Grade 3 PATs is the weakest followed support for the Grade 6 and 9 PATs, and the Grade 12 DEs.

When asked why they felt the way they did, the majority of comments made related to why the tests/examinations should not be administered, with only a few comments supporting the use of the tests/examinations. The two most frequently comments made concerned the inappropriate age of and stress placed on students in Grade 3, and, to a lesser extent, Grade 6, and the questionable validity of the scores in light of the failure to assess the full curriculum.

*Confidence about Knowledge of
Provincial, National, and International Testing Programs*

The teachers, principals, and superintendents were asked how confident they were with their knowledge about the provincial achievement testing and diploma examination programs, and the national and international (e.g., PISA, TIMMS) assessments administered in Canada. A five-point Likert scale (1= not confident, 5 = very confident) was used for this purpose. The mean and standard deviation for each question are reported in Table 9. Clearly, the teachers and principals were confident with their knowledge about the tests or examinations that were administered in their schools. The mean confidence levels varied between 4.51 and 4.60. The Grades 3 and 6 teachers and the PAT principals were not confident about their knowledge about the DEs. In contrast the Grade 9 teachers were confident about their knowledge of the DEs, the Grade 12 teachers and were somewhat confident with their knowledge of the PATs, and the DE principals were confident about their knowledge of the PATs. Lastly, the teachers and principals at all grades were not confident about their knowledge of national and international tests. In contrast, the superintendents were confident about their knowledge of the PATs, DEs, and national and international tests (Note 6).

Conclusions

Large-scale assessment programs are instruments of public policy (Mazzeo, 2001). The essence of educational assessment is making judgments about what students know or can do and, increasingly, how schools perform (Delandshere, 2001; Ryan,

2002). Recently, their use has expanded to include system accountability (Firestone, Mayrowetz, & Fairman, 1998). While there is a research tradition examining the technical concerns and issues of large-scale assessments, the acceptability of the role of these assessments in education has not been as carefully examined.

The results of the present study indicate clearly that the appropriateness of the purposes/uses of and seriousness of the issues raised about the provincial testing program are of concern to teachers, principals, and, perhaps not to the same degree, superintendents. Taken together, the actual uses of the results and, by implication, not mentioned non-uses of the results, correspond to the ratings of appropriateness of the purposes/uses and the seriousness of issues provided by the teachers, principals, and superintendents. The actual uses correspond closely to the purposes/uses initially set for the PATs and DEs. The PAT and DE results are less used for the additional purposes such as those related to accountability and the measurement of change. As one teacher succinctly put it:

I do think we need province-wide testing, but I think we need to look carefully at how and why we do them, and what uses are made of the results. What really irks me is the school to school comparisons that are made using the results from the test – so that upper middle class schools, full of very advantaged students, are shown to surpass some of our inner city schools, where caring teachers are burning themselves out working with students with a wide range of problems. It's not fair -- to any of us.

A superintendent added:

Alberta Education overemphasizes one-a-year assessments and is obsessed with accountability. Current assessment trends, research, and literature are promoting formative assess for learning, which promotes better, more

effective learning by students and improves student confidence in their ability to learn.

In connection with this latter comment, students are administered approximately 600 assessment instruments procedures, the results from which are included in student report cards (Rogers, 1991). The externally mandated provincial tests and examinations number about 16, and come at the end of the year or semester, with the results available approximately six weeks later. Thus, it is not surprising to see the low ratings associated with *assessment for learning* and *identify students in need of special services*.

The results of his study have credence with the finding, albeit with self-report data, that the teachers, principals, and superintendents expressed high levels of confidence in their knowledge about the tests and examinations that were administered to their students and less confidence in their knowledge about national and international tests that are administered to samples of students and not all students in the province. Further, as suggested in the teacher's comment, there is no overwhelming call of the elimination or reduction in the number of tests/examinations administered, except at Grade 3.

Have the provincial testing programs in Alberta promised too much? The answer to this question is yes. The purposes of the provincial tests have been expanded, but to the detriment of the acceptability of the tests. It is our considered opinion that while rated the highest, the unexpectedly low ratings of appropriateness of the purposes/uses associated with improving and enhancing teaching, instruction, and curriculum is attributable to the deep routed feelings and beliefs about the purposes added to the testing program and the use of the results to rank schools. It should be noted that no serious changes, other than making the tests secure, were made to the testing processes. Added purposes related to accountability – personnel and program evaluation and publicly reported school rankings – “spilled over” and influenced negatively the other appropriateness ratings. The provincial government needs to revisit the purposes they have set for these tests. Then the question needs to be asked: Is the testing program, in its present form, being asked to do

too much? As indicated above, since its inception, other than increasing the number of tests administered, no changes have been made, and no additional data are being collected in Alberta.

Other large-scale testing programs include much more data than presently collected in the large-scale testing program in Alberta. For example, student, teacher, and principal questionnaires are used to collect context, input, and process information that can be used to better understand achievement results. Context variables describe the economic and social factors believed to have an effect on the academic performance of students, classes, and schools and that are beyond the direct control of the educational system. Input variables include the resources that go into the system and that are open to the control. Process variables include the activities that are ongoing during the school year and are open to control. Collection of this information can be used in two complementary ways. Schools with like characteristics can be identified and placed in a group. The full set of results – context, input, process, and outcomes – is then shared among these similar schools. A principal and the teachers can be empowered to make changes by looking the “schools like me” to see how they are performing and what might be done to make improvements. Second, statistical procedures like hierarchical linear modeling (Raudenbush & Bryk, 2002; Snijders & Bosker, 2000) can be employed to see what variables impact performance at the student , class, and school levels (e.g., see Klinger, Rogers, et al., 2006)., Rogers, Ma, Klinger, et al., 2006). Together, the results of these two approaches will provide information should empower principals and their staffs to see ways that they can improve the performance of their students and schools in a sound and responsible way.

Notes

- ¹ Copies of the questionnaires are available from the first author.
- ² After the data collection had started, four of the larger school districts informed the researchers that we had to work through the districts research office. This led to the extended period required to complete the data collection from superintendents and principals.
- ³ The number of principals and superintendents are lower than expected. While the Executive of the College of Alberta School Superintendents endorsed the study and encouraged the superintendents to respond and distribute the Principal Questionnaire to all or some of their principals, only 23 (37.1%) superintendents complied with this request.
- ⁴ A private organization, using data obtained under the Access to Information Act, publicly ranks schools that are published in local newspapers. Often the rankings reported are for the schools in the major market for each paper, and not for the full province.
- ⁵ They were also asked if there were other grades at which the testing should be conducted. The number of respondents who said yes was small (less than 1%) and there was no discernible pattern in their listings of the other grades. Consequently, these results are not reported.
- ⁶ Three one-way ANOVAs were performed. The first analysis included the responses of the teachers at the corresponding PAT level, the responses of the Grade 12 teachers for the DEs, the responses of the principals with PAT grades, the principals with Grade 12, and the superintendents. The responses included in the second analysis were for the mirror image of the tests in the first analysis (e.g. Grade 3 teachers confidence of their knowledge about DEs). The third analyses included the responses of all groups about how confident they were about their knowledge of national and international tests. A significant difference was found in all three cases. The power was close to one for each analysis. Given the small value of the mean square residuals, the power was close to one. The differences, when they occurred, were not that meaningful despite their large

effect sizes. For example, in the first analysis, the level of confidence about the PATs reported by the Grade 9 teachers, 4.28, was found to be significantly lower than the confidence levels indicated by all the other groups regarding their knowledge about their grade and school. Yet, the Grade 9 teachers expressed a high level of confidence. In the case of the second analysis, the following results were found: $(3 = 6) < PAT_{Pr} < 12 < 9 < DE_{Pr} = S$). However, inspection of the means for this result reveals that the Grades 3, 6, and 12 teachers and the PAT principals did not express a high degree of confidence. Lastly, the third analysis yielded $(3 = 6 = 9) < (12 = PAT_{Pr}) < DE_{Pr} < S$. However, except for the mean of the superintendents, the means for the teachers and principals suggested that they were not confident about their knowledge of national and international tests. Consequently, the statistical results were disregarded.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, J.O., Muir, W., Bateson, D. J., Blackmore, D., & Rogers, W. T. (1990). *The impact of provincial examinations on education in British Columbia: General report*. Report submitted to the British Columbia Ministry of Education.

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1994). Policy and validity prospects for performance based assessment. *Journal for the Education of the Gifted. Special Issue: Reformers speak*, 17, 331-353.

Broadfoot, P. M. (1996). *Education, assessment and society*. Buckingham, UK: Open University Press.

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1994). Policy and validity prospects for performance based assessment. *Journal for the Education of the Gifted. Special Issue: Reformers Speak*, 17, 331-353.

Cheliminsky, E., & York, R. L. (1994). *Educational testing: The Canadian experience with standards, examinations, and assessments*. General Accounting Office Report PEMD-93-11. Gaithersburg: MD. GAO.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Delandshere, G. (2001). Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education*, 8, 113-133.

Dorn, S. (1998). The political legacy of school accountability systems. *Educational policy Analysis Archives*, 6. [<http://olam.ed.asu.edu/epaa/v6n1.html>].

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: The MIT Press.

Firestone, W., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, 20, (2), 95-113.

Hanson, F. A. (1993). *Testing, testing: social consequences of the examined life*. Berkeley, CA: University of California Press.

Klinger, D. A., Rogers, W. T., Anderson, J. O., Poth, C., & Calman, R. (2006). Contextual and school factors associated with achievement on a high-stakes examination. *Canadian Journal of Education*, 29 (3), 1-28.

Kvale, S. (1993). Examinations reexamined: Certification of students or certification of knowledge? In S. Chaiklin and J. Lave (Eds.), *Understanding practice: Perspectives on activity and context*. New York: Cambridge University Press.

Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.), *Educational Measurement, Third edition* (pp. xx). New York: Macmillan.

McLean, L. D. (1985). *The craft of student evaluation in Canada*. Toronto, ONT: Canadian Education Association.

Principles for Fair Student Assessment Practices for Education in Canada. (1993).Edmonton, AB: Joint Advisory Committee.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and Data Analysis* (2nd ed.). Thousand Oaks, CA. Sage Publications Ltd.

Rogers, W. T., Ma, X., Klinger, D. A., Dawber, T., Hellesten, L., Nowicki, D., & Tomkowicz, J. (2006). Examination of the influence of selected factors on performance on Alberta Learning achievement tests. *Canadian Journal of Education*, 29 (3), 708-733.

Rogers, W. T. (1991). Educational measurement in Canada: Evolution or extinction? *Alberta Journal of Educational Research*, 37, 179-192.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership*, 46, 4-9.

Shepard, L. A. (1991). Psychometrician's beliefs about learning. *Educational Researcher*, 20 (6), 2-16.

Snijders, , T., & Bosker, R. (2000). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publication Ltd.

Traub, R. E. (1990). Assessment in the classroom: What is the role of research. *Alberta Journal of Educational Research*, 36, 85-91.

Wilbrink, B. (1997). Assessment in historical perspective. *Studies in Educational Evaluation*, 23 (1), 31-48.

Table 1
<i>Purpose/Uses and Issues: Alberta</i>
<p>1. What are the purposes and uses of the Provincial Achievement Tests (PATs) administered at Grade 3 (Language Arts and Mathematics) and Grades 6 and 9 (Language Arts, Mathematics, Science, and Social Studies)?</p> <ul style="list-style-type: none"> ➤ Improve the achievement of all students while narrowing the gap among students ➤ Improve and enhance teaching and instruction ➤ PATS are one valid, reliable large scale measure that is part of the total continuous assessment of students ➤ Address issues of fairness; a common yardstick ➤ Communicate openly with parents/guardians about the achievement of their child(ren) and with the public at large ➤ Help students know what they are learning and to reflect on their progress and improvement ➤ Important part of a large scale accountability structure (see below, Accountability Pillar) <ul style="list-style-type: none"> • Part of the business plan for education • Look at percent who are deemed acceptable or excellent • Considered hard data • Engine that drives change <p>Diploma Examinations for each Grade 12 course (e.g., English 30 and 33, Pure Mathematics 30 and Applied Mathematics 30)</p> <ul style="list-style-type: none"> ➤ Certify the competence of students ➤ Selection for tertiary education together with post-secondary placement (degree and program) ➤ Improving teacher productivity ➤ Bring balance to the marking system: 50% teacher awarded mark and 50% examination mark ➤ Motivate students to work hard through Grade 12 ➤ Ensure fairness across the province ➤ Monitor the curriculum ➤ Important part of a large scale accountability structure <ul style="list-style-type: none"> • Part of the business plan for education • Look at percent who are deemed acceptable or excellent • Considered hard data • Engine that drives change

Table 1 (Cont.)

2. What issues do you see today with respect to the

PATs

- Grade 3: some people are concerned that this is too young at age at which to introduce provincial testing
- The participation rate at Grade 9 shows a negative trend
- How to assess special students, including IOP students? Should they be administered the PATs?
- Ineffective use of the provincial test results by teachers to improve teaching and instruction
- Ranking of schools based on provincial test results without taking into account context factors (e.g., simple Fraser Institute rankings)
- Failure to take into account context when interpreting school level results
- Some teachers move from or will not teach Grades 3, 6, and 9 to avoid provincial examinations. Some “not-so-good” teachers moved by principal.
- Incorrect credit given say to a Grade 3 teacher but not to the Grade 1 and 2 teachers given Grade 3 performance is cumulative (Grade 1 to 3).
- Pre-service teacher education is weak in the area of measurement and evaluation
- Lack of continuing education in measurement and evaluation

Diploma Examinations?

- Ineffective use of the diploma examination results by teachers to improve teaching and instruction
- Ranking of schools based on provincial test results without taking into account context factors (e.g., simple Fraser Institute rankings)
- Failure to take into account context when interpreting school level results
- Students erroneously see diploma examinations in English and Mathematics as screening tests
- University acceptance of all examinable courses (e.g., Applied Mathematics 30 not accepted by all; NAIT did and has now said no)
- Teachers who write items and take part in scoring are looked upon suspiciously by other teachers (participation violates union perspective)
- Alberta students are treated unfairly by tertiary institutions in provinces without provincial testing that counts toward final blended mark. The students are disadvantaged at the selection stage.

Table 2

		Teachers				Principals		
		3	6	9	12	PATs	DEs	Super.
Gender	Female	78.8%	59.5%	52.1%	42.2%	50.8%	23.8%	13.0%
	Male	21.2	40.5	47.9	57.8	49.2	76.2	87.0
Highest Level of Education	Certificate	1.7	2.7	0.4	0.6	0.8	-	-
	B.Ed.	56.3	57.3	40.7	27.2	17.8	9.5	-
	Degree + B.Ed.	31.9	27.3	37.8	54.5	11.0	19.0	-
Education	M.A./M.Ed.	10.1	10.7	14.6	15.5	66.9	61.9	78.3
	Ph.D./Ed.D.	0.0	0.7	0.4	0.2	2.5	9.5	21.7

Table 3

<i>Description of School</i>					
<i>Average family income</i>					
Grade	Far above	Above	Average	Below	Far below
Marked	average	average		average	average
3	7 (5.8)	17 (14.0)	45 (36.9)	38 (31.4)	14 (11.6)
6	12 (8.2)	24 (16.3)	56 (38.1)	41 (27.9)	14 (9.5)
9	13 (5.3)	74 (30.3)	102 (41.8)	48 (19.7)	7 (2.9)
12	47 (9.9)	151 (31.8)	185 (38.9)	70 (14.7)	18 (3.8)
Pr - PATs	3 (2.4)	20 (15.9)	52 (41.3)	7 (26.2)	18 (14.3)
Pr - DEs	2 (5.9)	4 (11.8)	16 (47.1)	7 (20.6)	5 (14.7)
<i>Academic Achievement of School</i>					
Grade	Far above	Above	Average	Below	Far below
Marked	average	average		average	average
3	6 (5.0)	37 (30.6)	49 (40.5)	24 (19.8)	5 (4.1)
6	6 (4.0)	36 (24.0)	74 (49.3)	28 (18.7)	6 (4.0)
9	15 (6.1)	92 (37.9)	106 (43.6)	28 (11.5)	2 (0.8)
12	30 (6.3)	171 (36.0)	212 (44.6)	50 (10.5)	8 (1.7)
Pr - PATs	7 (5.6)	44 (35.2)	51 (40.8)	17 (13.6)	6 (4.8)
Pr - DEs	3 (8.8)	11 (32.4)	13 (35.2)	3 (8.8)	4 (11.8)
Super. ^a	2 (8.7)	12 (52.2)	9 (39.1)	-	-
<i>Location of School</i>					
Group	Inner	Urban	Suburban	Semi-	Rural
	City			rural	
3	11 (9.2)	49 (40.8)	19 (15.8)	14 (11.7)	27 (22.5)
6	11 (7.4)	47 (31.5)	29 (19.5)	29 (19.5)	33 (22.1)
9	9 (3.7)	63 (26.0)	25 (10.3)	61 (25.2)	84 (34.7)
12	33 (6.9)	191 (40.2)	78 (16.4)	98 (20.6)	70 (14.7)
Pr - Pats	19 (15.1)	47 (37.3)	12 (9.5)	19 (15.1)	29 (23.0)
Pr - DEs	3 (8.8)	15 (44.1)	2 (5.9)	6 (17.6)	8 (23.5)

^a The superintendents rated the performance of their districts relative to other districts.

Table 4
*Appropriateness of Purposes/Uses
of Provincial Achievement Tests*

Group	\bar{Y}	s_Y	α	MS_{res}	F	Diff	ES
<i>Assessment for Learning k = 4 (20)</i>							
T	10.15	3.43	0.78	11.42	7.07*	LoT ^a	-
P	9.15	3.12	0.78				
S	11.48	3.63	0.85				
<i>Assessment of Learning k = 6 (30)</i>							
T	15.25	5.37	0.85	26.49	2.51	-	-
P	14.20	4.38	0.74				
S	15.88	5.70	0.68				
<i>Improvement of Curriculum and Instruction k = 5 (25)</i>							
T	16.01	4.15	0.80	16.60	14.76*	(T=P)<S	0.49
P	14.20	3.89	0.78				0.76
S	18.01	3.26	0.77				
<i>Provide Data for Data-based for Decision-Making k = 5 (25)</i>							
T	12.85	5.20	0.94	21.61	8.45*	(T=P)<S	0.84
P	13.51	4.79	0.89				0.70
S	16.78	4.43	0.84				
<i>Provide Information for Personnel and Program Evaluation k = 8 (40)</i>							
T	16.00	7.44	0.96	53.15	2.90	-	-
P	15.22	6.71	0.92				
S	19.17	6.98	0.93				
<i>Identify Student and School Needs k = 2 (10)</i>							
T	4.90	2.32	0.85	5.10	6.47*	LoT	
P	4.19	1.98	0.76				
S	5.52	2.33	0.87				
<i>Assessment of Progress of Cohorts k = 3 (15)</i>							
T	7.60	3.22	0.91	10.11	4.72	(T=P)<S	0.45
P	7.52	3.08	0.88				0.48
S	9.04	2.76	0.86				
<i>Provide Information for Parents and the General Public k= 3 (15)</i>							
T	7.06	3.05	0.87	9.18	5.61*	(T=P)<S	0.70
P	7.37	2.93	0.84				0.59
S	9.17	3.13	0.86				

Notes:- T: Teachers (n = 518; P: Principals (n = 129); Superintendents (n = 23)

*p < 0.05

k =x (y): the number of items and maximum possible score.

^a LoT: Lack of Transitivity

Table 5
*Appropriateness of Purposes/Uses
of Diploma Examinations*

Group	\bar{Y}	s_Y	α	MS_{res}	F	Diff	ES
<i>Provide Information for Post-Secondary Education k = 2 (10)</i>							
T	6.64	2.07	0.85	4.29	0.97	-	-
P	6.64	2.27	0.94				
S	6.26	2.00	0.85				
<i>Assessment of Learning k = 7 (35)</i>							
T	23.56	5.96	0.89	34.94	19.49	P<(T=S)	0.45
P	20.89	5.70	0.85				0.42
S	23.39	5.97	0.89				
<i>Improvement of Curriculum and Instruction k = 5 (25)</i>							
T	17.21	4.24	0.82	17.73	5.68	P<(T=S)	0.59
P	14.74	4.04	0.78				0.63
S	17.39	3.82	0.82				
<i>Provide Data for Data-based Decision-Making k = 5 (25)</i>							
T	14.37	4.55	0.91	20.94	1.21	-	-
P	14.77	4.52	0.93				
S	15.74	5.18	0.91				
<i>Provide Information for Personnel and Program Evaluation k = (40)</i>							
T	18.52	7.07	0.94	49.90	1.36	-	-
P	17.48	6.59	0.93				
S	20.59	7.62	0.96				
<i>Identify Student and School Needs k = 2 (10)</i>							
T	4.66	1.94	0.75	3.75	2.57	-	-
P	3.93	1.81	0.61				
S	4.90	2.03	0.73				

Notes:- T: Teachers (n = 475); P: Principals (n = 129); Superintendents (n = 23)

*p < 0.05

k = x (y): the number of items and maximum possible score.

^a LoT: Lack of Transitivity

Table 6
*Seriousness of the Issues Raised about
 Provincial Achievement Tests and Diploma Examinations*

Group	\bar{Y}	s_Y	α	MS_{res}	F	Diff	ES
<i>Characteristics of the PATs and DEs k = 7 (35)</i>							
T _{PAT}	21.19	6.25	0.80	36.73	1.65	-	-
T _{DE}	18.73	5.59	0.77	30.74	20.14	LoT ^a	
P	22.19	5.29	0.67				
S	20.50	5.79	0.78				
<i>Educators' Knowledge of How to Interpret and Use Results k = 6 (30)</i>							
T _{PAT}	16.20	6.73	0.94	41.46	22.93	(P=S)<T	0.63, 0.59 ^b
T _{DE}	16.68	6.58	0.93	39.79	4.87	(P=S)<T	0.74, 0.70
P	12.14	5.34	0.94				
S	12.39	5.47	0.95				
<i>Parent, General Public, and Press Knowledge k = 3 (15)</i>							
T _{PAT}	13.19	2.13	0.71	4.61	3.85	-	-
T _{DE}	12.81	2.40	0.82	4.56	1.59	-	-
P	12.72	2.20	0.63				
S	12.00	2.26	0.59				
<i>Use of Results for Accountability k = 3 (15)</i>							
T _{PAT}	13.02	2.46	0.76	6.26	9.02	S<(T=P)	0.87, 0.71
T _{DE}	12.41	2.63	0.79	6.94	4.35	S<(T=P)	0.59, 0.68
P	12.63	2.61	0.63				
S	10.85	2.26	0.59				
<i>Negative Impact on Curriculum and Instruction k = 7 (35)</i>							
T _{Gr3}	23.19	5.62	0.78	34.41	13.84	S<(T=P)	1.12
P	24.03	5.59	0.81				1.26
S	16.64	5.55	0.85				
T _{Gr6}	24.53	5.97	0.82	34.25	18.29	S<(T=P)	1.35
P	24.03	5.59	0.81				1.26
S	16.64	5.55	0.85				
T _{Gr9}	23.35	6.18	0.84	36.53	12.18	S<(T=P)	1.12
P	23.30	5.34	0.80				1.11
S	16.61	5.52	0.84				
T _{Gr12}	23.12	6.38	0.84	37.56	16.65	S<(T=P)	1.23
P _{Gr12}	22.79	6.57	0.89				1.18
S	15.57	5.17	0.81				

Notes:- *p < 0.05

k = x (y): the number of items and maximum possible score.

^a LoT: Lack of Transitivity

^b Effect Size, Teacher vs. Principals, Effect Size, Teachers vs. Superintendents

Table 7
Reporting Levels

Results	Report to .. (%)				
	Students	Parents	Teachers	Principals	Public
Student	68.4, 91.1, 96.0, 77.7, 87.0*	79.4, 90.2, 93.3, 83.5, 95.7*	82.7, 90.7, 95.3, 84.2, 100.0*	76.1, 80.9, 80.0, 82.0, 100.0	2.9, 7.3, 6.3, 7.9, 8.7
Class	14.0, 31.3, 35.6, 18.7, 30.4*	17.3, 30.5, 32.6, 19.4, 30.4*	81.2, 92.3, 95.8, 83.5, 95.2*	76.9, 84.1, 90.3, 82.7, 95.7	2.6, 5.3, 5.9, 2.2, 0.0
School	14.8, 28.9, 36.8, 34.5, 52.2*	31.6, 40.7, 46.1, 59.0, 69.6*	74.8, 87.8, 90.3, 85.6, 87.0*	82.4, 92.3., 94.7, 85.6, 100.0*	14.4, 28.6, 32.6, 34.5, 52.2*
District	12.9, 30.1, 35.2, 31.7, 34.8*	32.8, 42.8, 46.7, 56.8, 78.3*	68.8, 79.3, 79.2, 73.4, 78.3*	76.5, 86.6, 89.1, 80.6, 93.7*	34.9, 55.7, 60.4, 62.6, 91.3*
Provincial	15.4, 41.9, 49.3, 38.1, 43.5*	39.0, 57.6, 56.6, 59.0, 82.6*	69.1, 78.0, 83.4, 74.1, 78.3	73.2, 80.9, 86.1, 77.0, 91.3	50.0, 69.1, 80.0, 77.0, 91.3*

Notes:- a, b, c, d, e corresponds, respectively, to the Grade 3 and 6 teachers combined, Grade 9 teachers, Grade 12 teachers, PAT and DE principals combined, and superintendents.

* $p < 0.05$, but lack of transitivity.

Table 8
Grades at which Province-wide Testing should be Conducted

Group	Grade				χ^2	Diff
	3	6	9	12		
T	32.8%	53.3%	69.1%	90.6%	52.82*	3<6<9<12
P	29.5	57.6	64.0	86.1	122.24*	3<6=9<12
S	52.2	69.6	87.0	100.0	13.02*	LoT ^a

Notes:- Teacher responses correspond to grades taught.

* $p < 0.05$.

^a LoT: Lack of Transitivity

Table 9
*Confidence about Knowledge of Provincial,
 National, and International Testing Programs*

Testing Program	Teachers				Principals		Superintendent
	3	6	9	12	PAT	DE	
PATs	4.51 (0.63)	4.51 (0.63)	4.28 (0.81)	3.46 (1.35)	4.59 (0.58)	4.58 (0.61)	4.59 (0.59)
DEs	1.88 (1.12)	2.01 (1.11)	4.01 (1.30)	4.60 (0.61)	2.82 (1.44)	4.42 (0.79)	4.59 (0.50)
National & International	1.37 (0.71)	1.38 (.73)	1.66 (0.99)	1.88 (1.11)	1.94 (1.13)	2.61 (1.09)	4.14 (0.83)