COMPARING PERFORMANCE-BASED ACCOUNTABILITY MODELS: A CANADIAN EXAMPLE

Sonia Ben Jaafar & Lorna Earl
Ontario Institute for the Study of Education, University of Toronto

The intention of Performance-Based Accountability (PBA) policies is to foster school changes to enhance student learning and success. The influence of variation in these approaches, however, has not been empirically determined. This article employs a new conceptual framework to describe PBA models and compare them across contexts. We conducted a comparative analysis, finding that three kinds of PBA models exist in Canada. In this article, we consider the policy-level contextual differences coordinating large-scale, provincial, student testing and the use of results, using Canada as an example.

Key words: large-scale assessment, policy, using data for decision making, standards

L'intention des politiques de responsabilisation basée sur la performance (RBP) est de favoriser, au sein de l'école, des changements qui améliorent l'apprentissage et le succès des élèves. Cependant, l'influence de la variation dans ces approches n'a pas été déterminée de manière empirique. Dans cet article, les auteures présentent un nouveau cadre conceptuel pour décrire les modèles de RBP et les comparent dans divers contextes. Leur analyse comparative leur a permis de découvrir l'existence de trois types de modèles de RBP au Canada. Prenant le Canada comme exemple, les auteures se penchent ici sur les différences contextuelles au niveau des politiques quant à la coordination des épreuves communes provinciales et à l'utilisation des résultats.

Mots clés : épreuves communes, utilisation de données pour la prise de décisions, normes

---

The educational reforms of the past thirty years have employed large-scale student testing as a mechanism for educational accountability (Popham, 1999). When student performance on such tests acts as an indicator of school effectiveness and is used to hold schools accountable for results, the mechanism is specified to be performance-based accountability (Firestone, Mayrowetz, & Fairman, 1998; Fitz-Gibbon & Kochan, 2000). The intention of performance-based accountability (PBA) is to foster school change to enhance student learning and success. But, the relationship between the PBA approaches of central authority and school-level, accountability practices remains unclear (Linn, 2003). Understanding how educators respond to different PBA systems promises insight to reduce the gap between PBA policy intentions and school practices (Goertz & Duffy, 2001, 2003; Smith, 2003). Characterizing the policy differences is only one step towards this insight, and the purpose of this article.

The conceptual framework employed in this article highlights the multidimensionality of PBA policies. We used the five-dimensions from the framework to explore the policy differences in Canadian provincial jurisdictions and found differences emerged across the jurisdictional PBA models. Canadian PBA models do not include the overt consequences that are found in the USA and the UK, and the differences are small and subtle. The current literature does not foreground these smaller differences and, in so doing, implies they are too subtle for a salient impact on practice. We question that assumption and are engaged in a program of research to investigate its impact. This article describes the first step in that program of research – a comparative policy analysis of PBA for secondary schools in all Canadian provinces and territories.

This article is divided into several sections. In the first section, we define PBA systems and introduce the literature used to develop the conceptual framework. In the body of the paper, we detail the method of the study and present the findings in two parts, dimensional and holistic. Finally, in the conclusion we discuss the empirical findings for future Canadian-based research and the theoretical contribution of this study to the field of educational accountability.

PERFORMANCE-BASED ACCOUNTABILITY SYSTEMS

The standards movement has altered the purpose and the use of large-scale student testing.[1] Goertz and Duffy (2003) outline this change: "Policy makers are turning to data from large-scale assessments to make certification decisions about individual students, and to hold schools and school districts accountable for the performance and progress of their students" (p. 4). The performance for which schools are being held accountable is measured by standard student testing, and the mechanism of accountability attached to the results of those tests, originally structured to be a part of a system, is now a system unto itself (Ranson, 2003). Although educators and some public-interest groups have protested the imposition of the high stakes often attached to student testing (see Feldman, 2000; Kohn, 2001; Ontario Secondary Schools Teachers' Feder-ation, 2002; Rapp, 2001), it is a global trend (Carnoy & Loeb, 2002; Earl, 1995; Earl, Jantzi, Levin, & Torrance, 2000; Goertz & Duffy, 2001; Greg-ory & Clarke, 2003; Hodgkinson, 1995; McDonald, 2002; McEwen, 1995).

Elmore and Fuhrman (2001) suggest that expecting PBA systems to improve education is based on a misconceived assumption that these systems will promote compliance. They argue that PBA is intended to draw attention to academic performance so educators will improve teaching and learning, and school authorities will attend to capacity building to support school improvements. They go on to say that compliance is insufficient: how teachers and administrators understand what the results represent and mean serves as the key to sustained school improvement. It is imperative to understand the elements of PBA systems and their relationships because how educators use large-scale assessment results in their work is subject to how PBA is constructed and delineated by central authority.

Researchers have begun to examine the perspectives of teachers regarding the impact of state-testing programs, given different accountability models (Abrams, Pedulla, & Madaus, 2003). Although there is acknowledgement that different models cause variation in local practices, scholars have not conducted a systematic examination of the characteristics of these models to better detail the effect of their characteristic features. This study moves the field forward by

investigating the fea-tures of PBA systems at the provincial level and differentiating between how they interact to create a policy model. We characterize and identify common PBA models in Canada, which is a first step to understanding the extent of the relationship between how these models are opera-tionalized at the local level as school practices and those PBA policies.

MODELS OF PERFORMANCE-BASED ACCOUNTABILITY SYSTEMS

The multi-dimensional framework developed for this study reflects a more comprehensive characterization of PBA than currently found in the literature. Two recent national American studies (Abrams et al., 2002; Carnoy & Loeb, 2002) codified state PBA systems based on the stakes attached to the test results. These classification schemes, which use the severity of consequences attached to the results, reflect a prevalent association between consequences and accountability in current thinking on educational accountability (Pearson, Calfee, Walker Webb, & Fleischer, 2002; Stecher, 2002). Although the consequences attached to the results are commonly equated with the degree of answerability from stake-holders, other important predicating conditions position the conesquences within a given PBA model. The framework employed in this study considers these factors and characterizes PBA using five dimensions.

Five Dimensional Model of PBA

Armstrong (2002) states that accountability systems start with values and beliefs that, when turned into theory of action and then design principles, will help achieve their purposes (p. 2). The conceptual framework that we used for our study defines the design principles of five dimensions that Armstrong references to distinguish PBA models: (a) testing structure, (b) standard setting, (c) consequential use of data, (d) reporting, and (e) professional involvement. Although we present each dimension separately, their intimate relationship renders their distinction imprecise. Moreover, their relationships within a single PBA model characterizes the system.

*Testing Structure***.** Testing Structure (D1) consists of the scope, prevalence, and timing of the tests. The subjects and grades selected for

testing (Ryan, 2002), coupled with the timing of the tests and release of results, create the skeleton of the system. The subjects selected for testing highlight the priorities of the education system by highlighting which standards in the curriculum are measurement-worthy (Darling-Hammond, 1997; Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Smith, 1991). In addition to the grade level and the extent of student inclusion (i.e., all students or sample of students), the timing of the tests is significant.

Standard setting. Standard setting (D2) considers the purpose of the testing strategy as defined by accountability goals. The purpose of a testing program is first understood through understanding why standards are necessary for content and performance (Cizek, 1996). The wave of reforms have called for criterion-referenced tests based on content and performance standards (Fast & ASR SCASS, 2002; U.S. Department of Education, 2002) detailed in policy documents for educational practice (Darling-Hammond, 1997; Hamilton & Koretz, 2002). Content standards are generally set in terms of learning outcomes in curriculum documents. Alternatively, there is less agreement upon performance standards, per-haps because performance standards match the purpose and the nature of tests (Cooley, 1991) by determining the "frame of reference for inter-preting performance in education (absolute or relative, norm-referenced, criterion-referenced, or standards-based interpretation, or some combination)" (Ryan, 2002, p. 456). Educators' use of results depends on the standards because the alignment between the tests and norm or criterion-referenced standards will frame their interpretations (Hamilton & Koretz, 2002). For example, norm-referenced tests call for ranking stud-ents, schools, or districts in relation to peers or corresponding organ-izations, whereas criterion-referenced testing calls for judgments based on previously detailed expectations of acceptability.

*Consequential Use of Data***.** Consequential use of data (D3) considers the performance judgment related to the standards described in Dimension 2. Student performance is measured by statewide/provincial tests, and value judgments are used to translate those results into something meaningful, for example, achievement levels in standards-based systems, or ranking of students in a normative system. The use of data collected from the testing has been highlighted as the noteworthy

dimension in the literature because accountability is often tantamount to attaching stakes to test results (see Abrams et al., 2002; Abrams et al., 2003; American Educational Research Association, 2000; Carnoy & Loeb, 2002). Overt consequences have been attached to test results for students and schools (e.g., school reconstitution, awards, sanctions, grade promotion). There are other means in which results are linked to consequences intended to capitalize on "informal social pressures," such as public reporting of school results to attract students and student-linked funding (Hess, 2002, p. 70).

Reporting. Reporting (D4) methods are, in part, a function of the performance standards in conjunction with the testing structure. The decisions made prior to the implementation of the tests set parameters around the reporting of student achievement. The aforementioned dimensions mold the skeleton of the report. But there is possibility for variation, such as different comparisons and levels of the aggregation of results. Adding to the possibilities of reported results is the combination of achievement with other indicators (e.g., gender, race, type of school). The indicators selected for comparisons guide readers to seek out the differences between defined student groups and to draw conclusions regarding performance levels of given groups. Finally, the intended audience for the report is an important feature of this dimension.

Professional Involvement. Professional involvement (D5) reflects educator involvement, which is essential for school-level change (see Hargreaves, Earl, Moore, & Manning, 2001; Nye, Konstantopoulos, & Hedges, 2004; Wenglinski, 2002). The involvement of professionals is distinguished in all models of PBA, although the degree and type of involvement differs depending on the system. Irrespective of the kind of involvement being promoted, the rationale is to increase teacher understanding of student work related to the standards and the test specifications (Tienken & Wilson, 2001). Another component relating to educator use of the results is how educators make meaning from the results to improve educational, school-level practices.

This study, which operationalizes the five dimensions, defines PBA models in Canadian jurisdictions, exposing the critical and subtle differentiating characteristics in seemingly similar PBA models.

METHOD

We operatitonalized the five-dimensional framework to facilitate a qualitative, comparative policy analysis of PBA. We defined policies as any text articulating the intentions of the central authority to guide the actions of participants in the educational system (Bascia, Cumming, Datnow, Leithwood, & Livingstone, 2005; Pal, 1997). We purposefully sought out from ministries of Education and other official provincial sources texts meeting this description, addressing provincial assessment programs and activities related to the results of those assessments. We identified online and print documents such as guidelines, regulations, rules, policies, or procedures and collected them into a single N6 file (QSR International Pty Ltd, 2003) for 10 of the 13 Canadian jurisdictions. We excluded Prince Edward Island because it has no provincial testing system. Northwest Territories and Nunavut were excluded because there was no accessible information on their systems, and we received no response from a series of requests to their respective ministries of education for information on their PBA system. Our data set was 298 documents.

The data were coded in N6 using descriptors that were developed theoretically for each dimension and some that emerged from the data set. The descriptors were constructed as nodes and the dimensions were tree-nodes. For each dimension, the coded data were compiled by jurisdiction according to the descriptors. These N6 data compilations, placed in a text file, were used to characterize the dimension for the jurisdiction. Once all five dimensions were described for a single jurisdiction, we organized them into separate files for each dimension, subcategorized by jurisdiction. When the protocol for one jurisdiction was completed, another jurisdiction was started. This process was completed over the course of seven months (February 2004 to August 2004).

We then used the dimensional files to complete the comparative policy analysis in two stages. First, we used the descriptors to construct a comparative scaffold for each dimension. The data for each jurisdiction were examined using the dimension-specific comparative structure to facilitate a reliable comparison of relevant elements in each dimension. These data summaries reorganized the information and served as a

mechanism to compare jurisdictions within individual dimensions. Second, we conducted a holistic comparison considering the interaction and overlap of the dimensions. This secondary comparative policy analysis respected the natural relationships among the dimensions and allowed for rational groupings of the provinces by PBA model.

RESULTS

The descriptions of the elements used in the analysis are embedded in the dimensional comparisons. These five subsections are followed by the holistic comparative analysis of the patterns of interactions between the dimensions in the PBA models.

Dimensional Comparisons

In each dimension, we identified key elements as indicators of importance, and searched each dimensional data file separately for each jurisdiction for these key elements. We extracted and summarized them in a tabular format. We employed this protocol to reduce the data for individual jurisdictions and facilitate the comparative analysis. We restricted the criteria considered for each dimension and the data summaries for each jurisdiction to those policies relevant to secondary schools.

*Dimension 1: Testing Structure***.** We established testing structure (D1) through examining three elements: the number of grades tested, who takes the test, and the time lag between the administration of the test and the reporting of the results. In addition to a total of grades tested at the secondary level, we also noted which grades were tested. This information related to who takes the test, data that indicates the pervasiveness of the testing, for example, whether all students take the test or only a sample across the system. Finally the turnaround times for the results have implications for the response of the educators. Table 1 presents the results for each jurisdiction for D1.

Table 1: Summary Table for Testing Structure (D1)

Provinces & Territories	**Number of Grades tested Grades tested**	**Who takes the test?**	**Time between administration and reporting**
New Brunswick	4 grades are tested: Grades 8, 10, 11, and 12	All students in grades 8 and 10 All students in tested courses in grades 11 and 12	1 month for grade 8 2 months for grades 10, 11 and 12
Nova Scotia	3 grades are tested: Grades 8, 9, and 12	All students in grades 8 and 9 All students in tested courses in grade 12	4 months for grade 8 *1 month for grade 12
Ontario	2 grades are tested: Grades 9, and 10	All students in grade 9 and 10	3 months for grade 9 6 months for grade 10
Quebec	3 grades are tested: Grades SIII, SIV, and SV	All students in tested courses for SIII, SIV, and SV	6-8 weeks for individual reports
Saskatchewan	2 grades are tested: Grades 11, and 12	A sample of students in grade 11 Students in grade12 who were taught by an non-accredited teacher	Immediate
Yukon	2 grades are tested: Grades 9 and 12	All students in grade 9 All students in tested courses in grade 12	*1 month for grade 12
BC	2 grades are tested: Grades 10 and 12	All students in grade 10 All students in tested courses in grade 12	4 months for grade 10 1 month for grade 12
Manitoba	2 grades are tested: Grades S1 and S4	All students in S1 All students in tested courses in grade S4	1 month for all
Alberta	2 grades are tested: Grades 9 and 12	All students in grade 9 All students in tested courses in grade 12	2 months for grade 9 3 weeks for grade 12

Provinces & Territories	Number of Grades tested Grades tested	Who takes the test?	Time between administration and reporting
Newfoundland & Labrador	2 grades are tested: Grades 9 and 12	All students in grade 9 All students in tested courses in grade 12	3 months for grade 9 1 month for grade 12

*There was no information available or attainable for the turnaround time for the grade 9 test

The testing structures were similar in all jurisdictions because most of them tested students in grades 9 and 12. Saskatchewan and Quebec were the only two jurisdictions whose testing structure focused primarily on terminal courses, that is, courses that are the final courses students take prior to graduating. New Brunswick and Nova Scotia were unique with their more frequent test administration. In the case of the terminal tests, the turnaround time was short because the results were reported with the report card grade. In the case of the earlier grades, the turn-around time varied, with results generally made available when the students were promoted to the next grade level. Ontario, a unique case in its testing structure, was the only province that did not administer terminal tests. It also released the results months after the administration of the tests.

Dimension 2: Standard Setting. Standards setting (D2) was established by four elements: (a) the stated primary purpose of the testing system, (b) form of performance, (c) acceptable performance, and (d) the source of curriculum alignment. We derived the first element from explicit statements about the purpose of the tests. We associated form of performance with the purpose and type of tests in the system, providing the frame of reference for interpreting performance. For example, the form could be absolute or relative, norm-referenced, criterion-referenced, standards-based interpretation, or some combination (Ryan, 2002, p. 456). Finally, we determined acceptable performance in reference to the form. For example, if absolute levels were being used, then one of those levels served as the cut-off that constitutes acceptable performance. We considered the final element, alignment of test to curriculum, only in

terms of the curriculum being that of the jurisdiction or external. Table 2 presents the results for each jurisdiction for D2.

Table 2: Summary Table for Standard Setting (D2)

Provinces & Territories	Stated purpose	Form of Performance	Acceptable performance	Alignment to curriculum
Alberta	In grade 9: Educational improvement In grade 12: Student certification	There are two levels; acceptable or excellent	85% of students at acceptable and 15% at excellent	Own
Ontario	In grade 9: Accountability & improvement In grade10: Student credentialing	Grade 9: There are 4 levels Grade 10: There are two levels; pass or fail	Grade 9: Level 3 is the minimum Grade 10: Pass is needed	Own
Yukon	Public account of curriculum implementations	Grade 9: There are two levels; acceptable and excellent Grade 12: percentage of students passing (Use Alberta's criteria)	Grade 9: 50% Grade 12: 85% at acceptable and 15% at excellent (Alberta)	Other (WNCP, modified BC)
Manitoba	To certify student learning To monitor system quality	1-3 math 1-5 English	2 math (good understanding) 3 English (at level)	Own
Newfoundland & Labrador	In grade 9: School improvement In grade 12: Student credentialing	9: 1 (very limited) – 5 (outstanding) 12: percent correct	9: level 3 (adequate) 12: 50% pass	Own

Quebec	To ensure curriculum acquisition	There are not levels, it is based on the percent correct on the tests	60% is the minimum	Own
Saskatchewan	Student credentialing	There are multiple levels depending on the subject	Level 3 is the minimum	Own
British Columbia	In grade10: Educational improvement In grade12: Student credentialing	12: percent correct	10: benchmark 12: 40%	Own
New Brunswick	In the Anglophone system: To measure the effect of the system, improve it, and for student credentialing In the Francophone system: For program evaluation	*In the Anglophone system: Grade 8 and 12: There are five levels of performance Grade 9: There are two levels; pass or fail	In the Anglophone system: Grade 11: 60% is the minimum In the francophone system: Grade 12: 55% is the minimum	Own
Nova Scotia	In grades 8 and 9: To improve the curriculum and its implementation In grade 12: To judge the effectiveness of public education	Percent of questions answered correctly	Grade 12: 50% is the minimum	Other (APEF)

In all jurisdictions, the purposes of testing for the earlier secondary grades differed from terminal tests. At non-terminal grade levels, the main stated reasons for testing were school improvement, curriculum delivery, and public transparency; whereas terminal tests certified student learning, a form of student credentialing. The exceptional case was Ontario, whose system did not have terminal tests. But, the Ontario Secondary School Literacy Test, first administered in grade 10, certified students, attesting that they had acquired an acceptable standard of literacy to graduate. The other unique case was Saskatchewan, whose grade-12 tests were administered only for students of non-accredited grade-12 teachers.

Finally, an interesting point of distinction was the element of curriculum alignment. All the jurisdictions claimed that the tests were aligned to the curricula. Yukon and Nova Scotia relied on an external curricula, whereas the remainder of the jurisdictions referenced internal curricula.

Dimension 3: Consequential Use of Data. We constructed the consequential use of data as a numerical representation, summarizing the qualitative data that combined the degree and the locus of responsibility. There were five possible loci of responsibility (student, teacher, school, district, or province), and three possible degrees of consequences. A low degree occurred when indirect consequences that fell out from the use of the data were the only possibilities (e.g., school choice). A moderate degree occurred when there was an expectation of using the data in the intended policy, but no follow-through to ensure that the actions occurred (e.g., school plans). A high degree occurred when direct consequences were enforced using the results of the tests (e.g., including the results in students' final grades). We examined the data file for D3 for each jurisdiction to determine what degree of consequence was attached for each locus of responsibility, from student to province. When the degree of the uses of the data was established, we recorded a score for the degree per locus. A low degree was 1 point; moderate degree was 2 points; and high degree was 3 points. This combination yielded a maximum D3 score of 15 (a high degree for all loci of responsibility). When we recorded the scores in tabular form, it was evident that no variation occurred among jurisdictions with respect to

students; whereas the system score, the sum of the degree of consequence for teacher, school, district, and province, varied across the jurisdictions. Table 3 presents the results for each jurisdiction for D3.

Table 3: Summary Table for Consequential Use of Data (D3)

Provinces & Territories	**Student score**	**System score (teacher + school + district +province)**	**Overall score for degree of consequence (Student + system score)**
Alberta	3	8	11
Ontario	3	6	9
Yukon	3	6	9
Manitoba	3	6	9
Newfoundland & Labrador	3	6	9
Quebec	3	4	7
Saskatchewan	3	4	7
British Columbia	3	4	7
New Brunswick	Anglophone system: 3 Francophone system: 3	Anglophone system: 4 Francophone system: 0	Anglophone system: 7 Francophone system: 3
Nova Scotia	3	1	4

Because each province and territory constructed high stakes for students, no variation occurred. Variation was evident across the jurisictions when we considered system use of the results. New Brunswick's Francophone sector did not claim any use of the data, and Nova Scotia minimally required system-level use of the results. Most of the jurisdictions required some use of the results by teachers, schools, districts, and/or the province in some combination. Alberta claimed the most use of the results in its system.

Dimension 4: Reporting. We established the criteria for reporting (D4) using four elements: (a) the number of reporting levels, (b) the kinds of dis/aggregation of results, (c) the comparisons highlighted, and (d) the indicators included in the reports. There are a number of parts of the sysem interested in the results: the student, the school, the family of schools, the district, or the provincial level. The number of reports

generated for each level represents the number of reporting levels. In those reports, the remaining elements address possibilities for the content. The kinds of dis/aggregation of results represent how the results were aggregated and disaggregated to offer information. This element captures how the re-ults were aggregated in reports: whether the results were aggregated to the district level or disaggregated by gender or socio-economic status. How the results are reconstructed allows for different kinds of comparisons, but not all are highlighted in reports. The element of comparisons highlighted represents the comparative illustrations and texts in the reports that draw a direct group or temporal comparisons, for example, the literacy test results between girls and boys. Finally, the last indicator is included as recognition that the achievement results are often not presented without other information, or indicators are intended to facilitate the interpretation of the results. Table 4 illustrates the four elements per jurisdiction that were established from the data set for D4 reporting.

Table 4: Summary Table for Reporting (D4)

Provinces/ Territories	Number of levels of report	Results are dis/aggregated by:	Comparison	Indicators
Alberta	4 levels of reporting	School District	Year by year School to. province District to. Province	Participation rates
Ontario	4 levels of reporting	Gender ESL/EDL Program stream (academic and applied)	Board to province School to board School to province	Student learning environment survey
Yukon	4 levels of reporting	Classes (grade) Curriculum objective School First Nation	Year by Year Cohort trend	Student/teacher ratio Expenditure/student
Manitoba	1 level of reporting	Student	N/A	N/A

Newfoundland & Labrador	4 levels of reporting	Strand, task, item, topic District Gender Economic zone Program (e.g. honors)	School to district School to province	Survey data for school report Graduation rates over time
Quebec	4 levels of reporting	School Board Private school Public school Language of instruction Gender	Year by year Administrative region Education sector Gender Subject	Graduation rates by cohort
Saskatchewan	1 level of reporting	Dimension/strand Attitude of student Practices of students Gender	Courses taught by accredited to non accredited counterparts	OTL Graduation rate Course registration
British Columbia	Grade 10: 4 levels of reporting Grade 12: 2 levels of reporting	Grade 10: subject, gender, aboriginals, district, school, ESL, Frimm. Grade 12: subject, district, school	Schools Districts Year by year	Aboriginals
New Brunswick	Francophone system: 3 levels of reporting Anglophone system: 4 levels of reporting	Gender District School	In both systems: Year by year School to exam mark District to province School to province School to school District to district In Anglophone	Enrolment rate

			system only: Gender Language	
Nova Scotia	3 levels of reporting	School District Content area Gender	Board to board Board to province	N/A

All the jurisdictions reported the results, and most did so at multiple levels. Only Manitoba reported to the schools only. The reports mostly contained the data aggregated for the districts, schools, and province, and then disaggregated for gender, and other indicators. Although the reports varied in their content, they all used the results for comparative purposes. Different combinations of comparisons include items such as student performance over time, performance between different sectors, gender, or programs. Finally, other indicators were equally varied with no patterns detected in the details of the reports.

Dimension 5: Professional Involvement. We established professional involvement (D5) using a numerical summarizing system that combined the degree of involvement and the phases of testing. The degrees of involvment were low, moderate, and high, with a parallel analytical approach as employed in D3. A low degree occurred when a select few educators were involved, a moderate degree occurred when a larger subgroup of teachers was involved such as when many apply to mark tests, and finally a high degree of involvement occurred when all educators related to the testing were involved.

The phases of testing are standard setting, test construction, student preparation, administration, scoring, and results interpretation. We categorized the first three processes as the pre-testing phase, and the latter three as the testing/post-testing phase. We examined each data file in each jurisdiction for each phase of testing for any indication of a degree of involvement from professionals. We recorded the numerical representation of degree in a tabular form per phase and summed the scores by phase for each jurisdiction. This format illustrated a pattern between the pre-testing and post-testing phases. The phase differentiation helped expose the distribution of involvement among jurisdictions with similar or identical degrees of professional

involvement. Table 5 summarizes the professional involvement scores of the phases for each jurisdiction.

Table 5: Summary Table for Professional Involvement

Provinces & Territories	Pre-testing phases (standard setting + test construction + student preparation)	Post-testing phases (administration + scoring + results interpretation)	Overall Professional involvement Score
Alberta	5	7	12
Ontario	5	7	12
Yukon	3	7	10
Manitoba	4	8	12
Newfoundland & Labrador	5	6	11
Quebec	5	7	12
Saskatchewan	6	8	14
British Columbia	4	7	11
New Brunswick	5	7	12
Nova Scotia	4	6	10

Table 5 illustrates that the degree of professional involvement throughout the testing process differed minimally across the jurisdictions. Even when we examined the difference between the pre-testing and the testing/post-testing phases, the variation remained small. A pattern occurred between the two phase groupings: the pre-administration activities consistently engaged a lower degree of professional involve-ment from educators than the administration and post-administration activities. This trend likely reflects the centralized nature of the PBA systems, where the standard setting and test construction engaged the least involvement, and student preparation, test administration, and scoring engaged the widest distribution of professional involvement. Yukon and Manitoba are the two jurisdictions with the greatest difference between the two phases of professional involvement, con-trasting with Nova Scotia and Newfoundland and Labrador, which have the least involvement.

Holistic Comparison

We found the comparative analyses of the individual dimensions useful to understand the different characteristics of the PBA models. In this section, we compare the interactions of the dimensions in each jurisdiction. These relationships define PBA models and not the isolated dimensional differences. We begin by presenting the interaction of the consequential use of data with professional involvement. The relative degree of each of these dimensions is at the core of the influence of PBA in practice. Theoretically and empirically, these two dimensions have demonstrated a substantive influence. Earlier models of educational accountability have pivoted on these two constructs (Abrams et al., 2002; Carnoy, Elmore, & Siskin, 2003; Dorn, 1998; Linn, Baker, & Betebenner, 2002; Petrie, 1987). Additionally, the numerical construction summarizing the qualitative document analysis for these two dimensions facilitated the multidimensional comparison. The roles of testing structure (D1) and standard setting (D2) contributed to the comparison of the jurisdictional PBA models. Finally, we examined the results from the reporting (D4) comparison.

The relationship between the consequential use of data and professional involvement was a key consideration in the identification of distinct PBA models. We graphed the numerical representations of these dimensions to facilitate a comparison between the jurisdictions. This approach illustrated that a direct comparison of the two dimensions proved uninformative. The degree of professional involvement was consistently higher than the degree of consequential use of data in all PBA models. These results reflected the lack of variation in the consequential use of data for students in combination with a lack of refinement in looking at the overall score of professional involvement.

Given these initial results, we employed another approach to refine the comparison. The system level consequential use of data was used with the phase-grouped (pre- and post-administration phases) professional involvement scores. The three numerical representations were graphed to facilitate the comparison. We examined this data representation for patterns between the PBA models. We compared the system-level consequential use of data with the degree of professional involvement in each of the phase groupings. The relative positions of

these three constructs to one another exposed the emphases in each PBA model. This comparative element proved to be important to establish the distinction between the jurisdictional PBA models.

Three combinations surfaced in this analysis. The first was a combination in which system consequential use of data was emphasized more than either phases of professional involvement. Alberta was the only jurisdiction whose PBA model was in this category. The second combination occurred when both phases of professional involvement were emphasized more than the system consequential use of data. The jurisdictions whose PBA models were described by this second combination were Quebec, Saskatchewan, Nova Scotia, and New Brunswick. Finally, the third combination occurred when the system consequential use of data was emphasized less than professional involvement in the test/post test phase of the process, but emphasized more than professional involvement in the pre-testing phase of the process. This third com-bination described the PBA models in Ontario, Manitoba, Newfoundland and Labrador, and Yukon. The only province left in question given this categorization was British Columbia where the consequential use of data intersected with professional involvement in the pre-testing phase. The other dimensions were subsequently examined to help the categor-ization process.

We examined testing structure (D1) and standard setting (D2) in correspondence with these three groupings. No informative patterns occurred in either of these isolated dimensions (D1 & D2) because their structures were predominantly parallel across the jurisdictions. Patterns emerged only after we established the initial categories based on the D3/D5 relationship. The jurisdictions whose PBA models minimized the emphasis of the system consequential use of data relative to professional involvement were also those that tested students more frequently, with the exception of Saskatchewan. This PBA model in four jurisdictions (Quebec, Saskatchewan, New Brunswick, and Nova Scotia) promotes the use of test results to inform professional practice, suggesting that the frequent monitoring relative to the rest of the jurisdictions is intended to direct improvement efforts relative to professional involvement to support student achievement. Saskatchewan was an exception in this grouping because its PBA model de-emphasized the system's

consequential use of data relative to professional involvement, but students were not tested in this category as frequently as their counterparts. Saskatchewan tested students only in their terminal year. However, grade 11 and 12 students whose teachers were not accredited were the only ones tested. The unique purpose of this approach was to monitor the students' achievement for quality of program.

The jurisdictions that placed greater relative emphasis on the consequential use of data administered tests at two grade levels, early in high school, typically the first point of streaming for secondary students, and in grade 12 at the end of their schooling. Our examination of the testing structure resolved the grouping of British Columbia's PBA model. Brit-ish Columbia tests in grades 10 and 12. British Columbia's PBA model fits with that of Ontario, Manitoba, Newfoundland and Labrador, and Yukon. In this grouping, these provinces emphasize consequential use of data in the pre-testing phase more than professional involvement, but not more than the involvement in the post-testing phase, suggesting that the pre-testing phase work is removed from school personnel, but its use is imposed upon them. The most extreme case illustrating this structure is Yukon's PBA model, where this territory adopted the curriculum standards from British Columbia and the tests and performance standards from Alberta.

Finally, Alberta's model is unique although it shares some dimensional elements with the models of the other jurisdictions. In addition to emphasizing the consequential use of data more than professional involvement, Alberta sets its own system performance standards. Apart from the Yukon, which adopted Alberta's standards, it is the only jurisdiction that set acceptable performance standards for both the system and individual students. In Alberta, 85 per cent of students were expected to attain the "Acceptable" level of performance and 15 per cent were expected to attain the "Excellent" level. In all other cases, performance levels for individual students were the only performance standards stated in the PBA model. (See Figure 1.)

Category of PBA	Jurisdictions (grade levels tested)
Maximizing consequential use of data relative to professional involvement	Alberta (9 & 12)
Mid – consequential use of data relative to professional involvement	Ontario (9 &10); Manitoba (9 & 12); Newfoundland & Labrador (9 & 12); Yukon (9 & 12); British Columbia (10 & 12)
Minimizing consequential use of data relative to professional involvement	Quebec (10, 11, & 12); New Brunswick (8, 10, 11, & 12); Nova Scotia (8, 9, & 12)

Figure 1. PBA Categorization of Canadian Jurisdictions

We did not incorporate reporting (D4) in the results for the holistic comparison. The absence of a pattern in the comparative analysis of reporting as an individual dimension restricts its contribution for PBA model categorization. There are three possible reasons for this issue. First, reporting practices may not be relevant because they could have been subsumed by the consequential use of data dimension. More specifically, they may not be important except for their relation to the consequential use of data. The requirement for an improvement plan using the results from the reports generated by central authority was a common item in the jurisdictions. This practice suggests that the type of report structure may not be as important as the presence or absence of a report. Second, the construction of the reports may be inconsequential in the PBA models of Canadian jurisdictions because no severe consequences were attached to the results. Because all the models had mild consequences for the system, it may not be as important to pay attention to the structure of the reports as when a central authority imposes high stakes (e.g., sanctions or school reconstitution). Although the relationship between consequential use of data and reporting remains theoretically substantial, reporting may still be a separate dimension. This example leads to the third possibility, that the construct was inadequately developed for a comparative analysis. If the analytical construct for reporting were underdeveloped in the framework, it would

follow that the findings would be inconsequential in the comparative analysis. These possibilities suggest greater attention be given to the development of reporting to ascertain its value in PBA model characterization.

DISCUSSION

With this article, we present the first application of a five-dimensional framework as an operational tool for comparative analysis of PBA policies. Canada, a group of independent jurisdictions without a federal mandate, proved an interesting case. The provinces and territories formally share ideas and practices through voluntary membership in the Council of Ministers of Education and define their success partly through comparing achievement to their national counterparts. Ben Jaafar and Anderson (2007) identified trends in educational accountability across jurisdictional borders illustrating the commonalities and purported contextual individuality.

In this analysis of provincial policies, the distinguishing characteristic used to identify PBA model types was the relative importance of consequential use of data (D3) to the degree of professional involve-ment (D5). Earlier studies valued both these constructs, but our analysis considers the relational component enhancing the characterization of PBA models. Even with this relational component, testing structure (D1) and standard setting (D2) were necessary for sufficient detail to dis-tinguish models. Collectively, these four dimensions adequately described PBA models for comparison and categorization. The findings support the adaptability of the operational framework to different environments to examine PBA models. The conceptual framework facilitated an analysis at two levels: first, for an intensive examination of each dimension, and second, for the comparison of whole PBA models.

In Canada, each jurisdiction continues to invest substantial resources to develop and implement individual PBA systems. Each central authority claims its model improves student achievement and school practices. They make this claim in the absence of empirical evidence comparing the influence of different models on practice. This study shows that there are essentially three different types of PBA models

operating in Canada, which is the first step to investigate their influence on practice. The model categorization can be used to guide impact studies to examine the influence of the different PBA models on school-level practices. Researchers and practitioners can use dimensional and holistic PBA comparison to examine school practices to tease out the influence of PBA policies and systems. Only in conducting this kind of inquiry will insight into influential, appropriate, and practical PBA models be established.

The broader implication of this work is that researchers can use the framework to expose policy similarities and differences within complex systems, offering potential for within-jurisdiction, national, and international comparisons. Given the current global policy trends of increasing use of PBA systems, there is urgency in documenting the impact of these models at the school level. Tests are increasingly being administered across national and international borders (e.g., National Assessment of Educational Progress [NAEP], Trends in International Mathematics and Science Study [TIMSS], Programme for International Student Assessment [PISA], and School Achievement Indicators Program [SAIP], replaced by Pan-Canadian Assessment Program [PCAP]) and the test scores are being used to feed an inappropriate "Olympics of education" (Creemers, 2005). Although some efforts are being made to qualify and value regional differences, the relative readiness of a numerical database is seductive for comparative analyses.

This article offers consideration at the policy-level of contextual differences coordinating large-scale provincial student testing and the use of results. We contend that educators and policy makers should consider these differences when they investigate results and deliberate on employing large-scale provincial testing. Although the theoretical contribution of a comprehensive, conceptual, policy-level model is important to the scholarship on educational accountability, the significant practical and theoretical value of the findings in this study will only be realized when they are employed in follow-up impact studies.

NOTES

[1] The term large-scale student testing is specific to provincial or state-level testing for the purpose of this article.

REFERENCES

Abrams, L., Clarke, M., Pedulla, J., Ramos, M., Rhoades, K., & Shore, A. (2002, April). Accountability and the classroom: A multi-state analysis of the effects of state-mandated testing programs on teaching and learning. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice,* 42(1), 18-29.

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in Pre-K-12 education. *Educational Researcher*, 29(8), 24-25. Retrieved May 21, 2008, from http://www.aera.net/policyandprograms/?id=378

Armstrong, J. (2002). What is an accountability model? [Issue Paper]. Denver, CO: Education Commission of the States.

Bascia, N., Cumming, A., Datnow, A., Leithwood, K., & Livingstone, D. (2005). Introduction. *International handbook of educational policy: Part one.* Dordrecht, The Netherlands: Springer.

Ben Jaafar, S., & Anderson, S. (2007). Policy trends and tensions in accountability for educational management and services in Canada. *The Alberta Journal of Educational Research*, 53(2), 205-225.

Carnoy, M., Elmore, R. F., & Siskin, L. S. (Eds.). (2003). *The new accountability: High schools and high stakes testing*. New York: RoutledgeFalmer.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*. 24(4), 305-331.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice,* 15(1), 13-21.

Cooley, W. W. (1991). State-wide student assessment. *Educational Measurement: Issues and Practice,* 10(4), 3-6.

Creemers, B. (2005, January 2-5). The contribution of international studies on education. Paper presented at the International Congress for School Effectiveness and Improvement, Barcelona, Spain.

Darling-Hammond, L. (1997). Creating standards without standardization. In *The right to learn: A blueprint for creating schools that work* (pp. 210-260). San Francisco, CA: Jossey-Bass Publishers.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives,* 6(1). Retrieved May 18, 2008, from http://epaa.asu.edu/epaa/v6n1.html

Earl, L. (1995). Assessment and accountability in education in Ontario. *Canadian Journal of Education,* 20(1), 45-55. Retrieved May 18, 2008, from the Canadian Society for the Study of Education (CSSE) Web site: http://www.csse.ca/CJE/Articles/FullText/CJE20-1/CJE20-1-05Earl.pdf

Earl, L., Jantzi, D., Levin, B., & Torrance, N. (2000). *OISE/UT evaluation of the implementation of the national literacy and numeracy strategies. First annual report*. *Watching & learning*. Toronto: Ontario Institute for Studies in Education; Department for Education and Employment, London, England. Retrieved May 18, 2008, from the Education Resources Information Center (ERIC) Web site: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1a/b1/3c.pdf

Elmore, R. F., & Fuhrman, S. H. (2001). Holding schools accountable: Is it working? *Phi Delta Kappan,* 83(1), 67-72.

Fast, E. F., & ASR SCASS. (2002). A guide to effective accountability reporting. Unpublished manuscript, Washington, DC.

Feldman, S. (2000, July). *Uproar over testing*. Washington, DC: American Federation of Teachers (AFT). Retrieved May 18, 2008, from http://www.aft.org/presscenter/speeches-columns/wws/2000/0700.htm

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

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Education Policy Analysis Archives,* 8(35), Retrieved May 18, 2008, from http://epaa.asu.edu/epaa/v8n35/

Fitz-Gibbon, & Kochan, S. (2000). School effectiveness and educational indicators. In C. Teddie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 257-282). London, UK: Falmer Press.

Goertz, M. E., & Duffy, M. C. (2001). *Assessment and accountability across the 50 States. CPRE Policy Briefs*. Philadelphia, PA: Consortium for Policy Research in Education. Retrieved May 18, 2008, from the Education Resources Information Center (ERIC) Web site: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/5f/13.pdf

Goertz, M. E., & Duffy, M. C. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory into Practice,* 42(1), 4-11.

Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory into Practice,* 42(1), 66-74.

Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 13-49). Santa Monica, CA: RAND Education.

Hargreaves, A., Earl, L., Moore, S., & Manning, S. (2001). *Learning to change: Teaching beyond subjects and standards* (1st ed.). San Francisco, CA: Jossey-Bass.

Hess, F. M. (2002). Reform, resistance, . . . retreat? The predictable politics of accountability in Virginia. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 69-104). Washington, DC: Brookings Institution Press.

Hodgkinson, D. (1995). Accountability in education in British Columbia. *Canadian Journal of Education,* 20(1), 18-26. Retrieved May 18, 2008, from the Canadian Society for the Study of Education (CSSE) Web site: http://www.csse.ca/CJE/Articles/FullText/CJE20-1/CJE20-1-03Hodgkinson.pdf

Kohn, A. (2001, June 30). More testing is no answer. *USA Today*. A14

Linn, R. L. (2003). *Accountability: Responsibility and reasonable expectations. CSE report* . Los Angeles, CA: Center for the Study of Evaluation. Retrieved May 18, 2008, from the Education Resources Information Center (ERIC) Web site: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/6a/9d.pdf

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher,* 31(6), 3-16. Retrieved May 18, 2008, from http://www.aera.net/uploadedFiles/Journals_and_Publications/Journals/Educational_Researcher/3106/3106_Linn.pdf

McDonald, M. (2002). The perceived role of diploma examinations in Alberta, Canada. *The Journal of Educational Research,* 96(1), 21.

McEwen, N. (1995). Accountability in education in Canada. *Canadian Journal of Education,* 20(1), 17. Retrieved May 18, 2008, from the Canadian Society for the Study of Education (CSSE) Web site: http://www.csse.ca/CJE/Articles/FullText/CJE20-1/CJE20-1-02McEwen1.pdf

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis,* 26(3), 237-257. Retrieved May 21, 2008, from http://www.sesp.northwestern.edu/docs/publications/169468047044fcbd1360b55.pdf

Pal, L. (1997). *Beyond policy analysis: Public issue management in turbulent times*. Scarborough, ON: International Thomson Publishing Nelson.

Pearson, P. D., Calfee, R., Walker Webb, P. L., & Fleischer, S. (2002). *The role of performance-based assessments in large-scale accountability systems: Lessons learned from the inside*. Washington, DC: Council of Chief State School Officers. Retrieved May 21, 2008, from http://www.ccsso.org/content/pdfs/TILSACalfee.pdf

Petrie, H. G. (1987). Introduction to "Evaluation and Testing." *Educational Policy,* 1(2), 175-180.

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice,* 18(3), 13-17.

QSR International Pty Ltd. (2003). N6 (Version 6) [Qualitative software]. Victoria, Australia.

Ranson, S. (2003). Public accountability in the age of neo-liberal governance. *Journal of Education Policy,* 18(5), 459-480.

Rapp, D. (2001). Ohio teachers give tests an 'F'. *Rethinking Schools Online,* 15(4). Retrieved August 4, 2003, from http://www.rethinkingschools.org/archive/15_04/Ohio154.shtml

Ryan, K. (2002). Shaping educational accountability systems. *American Journal of Evaluation*, 23(4), 453-468.

Smith, M. (2003, April 21). Accountability in educational reform: Tensions and dilemmas. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

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

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79-100). Santa Monica, CA: RAND Education.

Tienken, C., & Wilson, M. (2001). Using state standards and tests to improve instruction. *Practical Assessment, Research and Evaluation*, 7(13). Retrieved May 21, 2008, from http://pareonline.net/getvn.asp?v=7&n=13

U.S. Department of Education. (2002). *No Child Left Behind Web Site*. Retrieved May 21, 2008, from http://www.ed.gov/nclb/index/az/index.html

Wenglinski, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, 10(12). Retrieved May 21, 2008, from http://epaa.asu.edu/epaa/v10n12/

*Sonia Ben Jaafa***r** is a recently graduated doctoral student from the Theory and Policy Studies Department at OISE/UT. She specializes in the how assessment and accountability can be used to improve educational systems. She is currently involved in supporting the major education reform in Qatar. In addition, as a Research Associate at Aporia Consulting Inc., she is co-authoring a book on Network Learning for Corwin Press. Her recent publications include Ben Jaafar, S. & Anderson, S. (2007). Policy trends and tensions in accountability for educational management and services in Canada, *Alberta Journal of Educational Research*, 53(4); Volante, L. & Ben Jaafar, S. (2008). Educational assessment in Canada. *Assessment in Education*; Ben Jaafar, S. (2007). Secondary school leaders respond to performance-based accountability in Canada, *Principal Matters*; Ross, J. & Ben Jaafar, S. (2006). Participatory needs assessment. *Canadian Journal of Program Evaluation*, 21(1), 131-154.

Lorna Earl is Director, Aporia Consulting Ltd. and a recently retired Associate Professor in the Theory and Policy Studies Department and Head of the International Centre for Educational Change at OISE/UT. Throughout her career, she has been involved in writing, research, consultation evaluation, and staff development with teachers' organizations, ministries of education, school boards, and charitable foundations in Canada, England, Australia, New Zealand, Europe, and the United States. Her recent publications include Earl, L. & Katz, S. (2006), *Leading in a Data Rich World*. Thousand Oaks, CA: Corwin Press; Earl, L. & Katz, S. (2006) *Rethinking classroom assessment with purpose in mind*. Western Northern Curriculum Partnership (WNCP); Earl, L. (2003) *Assessment as learning: Using classroom assessment to maximize student learning*, Thousand Oaks: Corwin Press.