Accountability Testing in Canada: Aligning Provincial Policy Objectives with Teacher Practices

Derek Copp
Good Spirit School Division

Abstract

One of the major functions of large-scale assessment is educational accountability. The expectation for the improvement of instruction based on test results is called "professional accountability" and it is built into provincial assessment policies across Canada. This study asked teachers to self-report on instructional changes they have made in response to large-scale assessment data. How teachers respond to external evaluation is "reactivity" which can gauge both the amount and type of instructional effects. Quantitative analyses were done to examine the prevalence of reactivity, and qualitative interview data were used to support the findings and to elaborate on relevant details. The results were that teachers generally do show reactivity to large-scale assessment data, yet this change tends to be more "teaching to the test." There is no consensus in the teaching community regarding assessment or accountability policies. These policies do not focus on teacher implementation practices needed to meet stated goals.

Keywords: educational accountability, professional accountability, large-scale assessment, reactivity, assessment policy, Canadian provincial assessment

Introduction

Modern efforts to measure the effectiveness of public sector services began with the introduction of the New Public Management (NPM) model in the United Kingdom in the 1980s. Accountability in this governance model emphasized transparency with public funds and functions (Shore & Wright, 2000; Van Thiel & Leeuw, 2002). Now, as then, large-scale assessment (LSA) is considered the most effective means to gauge the performance of the educational system. LSA provides ministries a measure to compare school divisions, schools, and individual teachers. The nearly ubiquitous current use of accountability metrics in schools does produce a lot of data, but it is less certain that accountability measures always help produce better schools (Espeland & Sauder, 2007).

Data collection (such as LSA scores) does not in itself indicate improvement or adherence to the stated aims of NPM reforms (Volante, 2007; Volante & Ben Jaafar, 2008). Data-driven systems need data, but they also should be used wisely. It is also the case that system-wide accountability goals are quite different from goals that seek to change an individual's behaviour. The distinction is made by Møller (2009): *political* and *public* accountability seems to drive educational assessment policies, and *professional accountability* relates more to teachers being devoted to professional improvement and acting as the public would expect them to act (putting students first, collaborating with colleagues, etc.). A focus on the former perspective is commonly thought to drain professional autonomy from educators and influence policy from an efficiency-based perspective. Fullan (2011) notes: "The focus on accountability uses standards, assessment, rewards and punishment as its core drivers. It assumes that educators will respond to these prods by

putting in the effort to make the necessary changes" (p. 8). It is problematic to assume that accountability measures set in policy by the education ministry will always work as expected at the classroom level with teachers (Volante, 2005; Wayman & Stringfield, 2006). Effective and consistent implementation is more than half the battle.

International LSAs and Public Accountability

The acceptance and use of large-scale international assessments is currently wide-spread as a means of gauging educational effectiveness, and the media give these results considerable attention (Benveniste, 2002; Morris, 2011). The most influential of these assessments is the Organization for Economic Cooperation and Development (OECD) "Program for International Student Assessment" (PISA). The 2015 edition was written in 70 countries, and while some nations take these results more seriously than others, an examination of documents from the ten provincial education ministries in Canada shows that all provinces are concerned with PISA rankings. This is partly because of how these are reported in the press (Breakspear, 2012; Fullan, 2011). For example: "The PISA . . . assessment rankings may place Saskatchewan students at a serious disadvantage for acceptance into post-secondary education programs of study, as well as employment opportunities" (Saskatchewan Ministry of Education, 2012, p. 3). PISA rankings often provide the justification and impetus for change to provincial testing systems. The same amount of emphasis is rarely placed on a small-scale but important policy goal: changes to classroom instructional practice. PISA data is also not generally considered useful at the classroom-level since no individual school's or teacher's data is disaggregated for this kind of utilization.

Provincial Testing and Professional Accountability

This accountability metric is focused on individual teachers and it is found in education ministry documents from all across Canada. Professional accountability goals found in the literature include (but are not limited to): (a) the improvement of student achievement (Hamilton, 2003), (b) the stimulation of professional reflection on current practices (Halverson & Thomas, 2007; Means, Padilla, DeBarger, & Bakia, 2009), and (c) both informing and initiating professional development (Goertz, Oláh, & Riggan, 2009; Johnson & Chrispeels, 2010). These are not simple tasks especially when a single assessment tool is used as a catalyst for all of them (Bolon, 2000).

The different levels of accountability written into provincial assessment programs are the primary reason that testing instruments are having their limits pushed in order to meet divergent ends. Table 1 shows how Canadian provincial jurisdictions have loaded their assessment policies with accountability goals ranging from the policy- to the classroom-level. This paper sets out to examine the classroom-level accountability goals of assessment policies and how they play out in the practice of teaching professionals.

Table 1
The Stated Policy Purposes of Provincial Assessment

Policy Level ⟸ ⟹ Classroom Level	**Purposes of large-scale provincial assessment**	AB	BC	MB	NB	NL	NS	ON	PEI	QC	SK
	Reporting/Accountability	x	*	x	x	x	x	x		x	*
	Requirement for graduation	♦	♦	♦	x	♦	♦	x		♦	●
	Reference to PISA, PCAP, PIRLS, TIMMS	x	x	x	x	x	x	x	x	x	x
	Monitor/improve student achievement	x	x	x	x	x	x	x	x	x	*
	Improve central data-based decision-making		*		x		x	x	*	x	*
	Adherence to curriculum	x			x	x	x	x	x	x	
	Interventions for struggling students	x	x	x	x			x	x	x	*
	Improve local data-based decision-making		*		x	x	x	x		x	*
	PD/assessment literacy of educators		x	x			x	*			
	Total number of stated/implied purposes	6	8	6	8	6	8	9	5	8	7

x - Policy evident from ministry literature
* - Policy not explicit, but apparent from ministry literature
♦ - Exams must be written but need not have a passing grade
● - Exams are mandatory when teachers are not accredited

Unintended Consequences

With this in mind, an examination of testing policy in Canada shows many of the same unintended consequences found in the literature from the more mature American system. The extra focus on tested subjects is known as "curriculum narrowing" and sees teachers neglect non-tested content in favour of that on LSAs (Koretz, 2002). Creating valid and reliable tests of deep understanding is both complex and expensive, and as a result many LSAs leave out outcomes that demand subjective marking or those that address complex outcomes (Ungerleider, 2006). Where tests are created with these considerations in mind, teachers often adjust their teaching to reflect what is on the test, both in terms of the content and the types of questions that are asked (Mintrop, 2003; Morris, 2011). Gauging how prevalent these practices are in Canada was one of the purposes of the research study.

This paper will be laid out in the following way: (a) the framework of reactivity will be explained including the researcher's unique reactivity model for data collection and analysis, (b) the methods used will be clarified, (c) the results from the national survey on reactivity effects will be discussed, (d) emergent themes related to reactivity and accountability will be examined through the transcripts from interviews, and finally (e) the conclusions will examine some of the implications of these findings.

Reactivity Model

The manner chosen to conduct this program evaluation study of large-scale assessment was using the framework of reactivity. First described by Donald T. Campbell (1957), reactivity describes how actors alter their behaviour when they know they are being externally evaluated. Campbell considered reactivity a flaw in social studies experimental research since such changes in behaviour give an inaccurate picture of the reality one is trying to uncover. At its most basic level, it means that actors in an audited system who are aware of the metrics being evaluated are often able and willing to highlight those qualities which will lead to the most favorable ratings. Reactivity effects are evident in institutions of higher education (Espeland & Sauder, 2007), therapeutic experimental studies (Matheson, Rogers, Kaskutas, & Dakos, 2002), and workplace behaviours (Lerner & Tetlock, 1999). Espeland and Sauder (2007) also make a clear distinction between legitimate professional activities and those undertaken only to improve scores on

these external metrics (see teaching to the test and teaching [to] the curriculum below). Since educational accountability data are collected for the purpose of making important policy decisions, it is important that rigorous analysis be conducted to confirm that data-based decisions are not in actuality "decision-based data collection."

Teaching to the Test

Espeland and Sauder (2007) looked at university law schools reacting to faculty rankings done by a US news magazine, and noted the actions they considered defensible as well as those they saw as "gaming behaviours." Instructional changes made in Canadian classrooms as reactions to provincial testing might also be categorized as either those that are legitimate and defensible, or those that are best seen as a means to improve test scores. The former category is defined in this paper as "teaching (to) the curriculum" (TTC). It includes those reactions to LSA testing which are both ethical and provide a wide range of outcomes for students. The items on this side of the list are generally accepted as pedagogically sound and are also supported by the professional codes of competence.

The alternative to teaching (to) the curriculum strategies are defined as teaching to the test (TTT) which includes those educational practices which are thought to be either unethical or reduce the number or variety of outcomes presented to students. These strategies are seen as the most direct methods of improving test scores, but they do not have the transferability upon which teaching (to) the curriculum practices are built (Au, 2007; Popham, 2001; Zigo, 2001). The distinction made here between TTT and TTC strategies was not made evident to survey respondents.

Data on reactivity were self-reported by teachers on a nationwide survey. The prompts found in table 2 were used in the research survey to determine types of reactivity employed in classrooms. Neither grouping was identified to the respondents. They were simply presented as possible instructional change reactions to LSA results. Respondents selected the five TTC and five TTT strategies as their manner of using LSA results. Neither list was intended to be fully exhaustive considering that all highly unethical options (such as outright cheating) were excluded. These are all, in appropriate circumstances, suitable preparation strategies. Yet TTT is not supported by ministry documents, the STF (2015) Code of Professional Competence, nor is it well-supported in the education literature especially when used by default or with whole-class groups. Koretz (2005) states the division line clearly:

> The problem of inappropriate test preparation has two related aspects. The first, already noted, is inflation of test scores. The second is undesirable pedagogy. This can take numerous forms, such as a boring drill and practice focused on test content or the elimination of important content not emphasized by the test. The two are obviously closely intertwined: undesirable forms of instruction are among the primary factors that cause inflation of scores. (p. 2)

A concern with self-reported data, as in this study, is that respondents may not honestly report socially undesirable actions (Ashton, Buhr, & Crocker, 1984). From the findings that follow it is clear that teachers in this study are unsure about which educational practices are best suited to high educational outcomes for students (Koretz & Jennings, 2010; Nagy, 2000). One can debate the relative merits of specific instructional methods, especially in terms of individual students. This divided list does not imply that any such classroom accommodations are unsound. In general, though, there is consensus that the following strategies do not best serve all learners in all cases: (a) narrowing the curriculum, (b) teaching skills that are useful only in specific testing situations, (c) restricting the use of varied and creative assessment methods, (d) side-lining non-core subjects, and (e) reviewing previous tests to prepare students for an assessment without developing any deep understanding of the content.

Table 2
Reactivity Prompts from the Data Gathering Survey

Q: Think about the ways your instruction may have changed in classes which write provincial assessments as compared to those classes that do not write these tests. Choose a response for all the following statements: (Response choices were: not at all; somewhat; a great deal.)	
Teaching (to) the Curriculum	**Teaching to the Test**
I have looked for Professional Development to improve my instructional practices.	I cover material I know will be on the test very well.
I have requested additional resources related to testing.	I focus more on test-taking strategies like the process of elimination.
I have worked with other teachers to make sense of the data.	I use the format of the test to give similar types of practice questions.
I cover a wider range of topics in the curriculum.	I focus more on subjects that have provincial tests.
I hold group study sessions or provide extra help after school.	I review old exam questions.

Take the illustrative hypothetical example of a large-scale mathematics test. A teacher could employ TTT strategies in several ways: (a) teaching in great detail concepts thought to be on the test; (b) teaching test-taking strategies not explicitly in the math curriculum; (c) doing practice math tests which use the same format as questions on the LSA; (d) excluding instruction of concepts not thought to be on the test; and (e) if questions or concepts are sometimes repeated, teaching problems from previous years' tests.

At the policy level, large-scale tests are intended to provide information about student performance for an entire curriculum, not simply a circumscribed domain from within that curriculum. As a consequence of limited time and money, they often do the latter. A key principle behind large-scale math testing is that you should be able to reasonably infer the general mathematics skills of students based on test results using more or less randomly selected items from the entire curriculum (again, not a small subset of concepts). That inference is made invalid when teachers emphasize (as test preparation) the concepts that are "less than randomly" selected. TTT practices may generally increase test scores, but in the process the purpose of the test is undercut, and the data provided to the ministry are no longer an accurate reflection of students' abilities.

At the classroom level, a student who is not particularly strong in mathematics may score well on such an LSA as a result of TTT strategies being employed. A serious consequence of this "score inflation" (Koretz, 2003) is that the classroom teacher may not accurately identify a student's weaknesses and therefore would dedicate no time or resources to their remediation. This subverts the potential for students to master important curricular outcomes.

Policy-Based Expectations

Canadian provincial education ministries have set out several goals to be met by their large-scale assessment programs (see table 3). The researcher identified nine goals and more than half of the provinces indicated that eight or more of these purposes were policy goals. The fewest identified goals by a single province was five. The following word-for-word quotations clarify both policy- and classroom-level expectations for data use which are integral to LSA policies:

> [Alberta:] The purpose of the Achievement Testing Program is to: determine if students are learning what they are expected to learn… Teachers and administrators can use this information in planning and delivering relevant and effective instruction in relation to learning outcomes in the Programs of Study. (Alberta Education, 2015)

> [British Columbia:] Using information from FSA, the Ministry of Education works with school districts to provide support for students and to improve teaching and learning for the coming school year… The BC Performance Standards are intended as a resource to support ongoing instruction and assessment. (British Columbia Ministry of Education, 2015)

> [Manitoba:] The Provincial Assessment Program supports learning by: providing feedback to students, teachers and parents about student learning; informing instructional planning and helping to determine the need for changes or student specific interventions… (Manitoba Education, 2015)
>
> [New Brunswick:] Assessment enables teachers to gather data to determine the needs of their students and to address those needs adequately in order to tailor instruction. Large-scale data gathered through the provincial assessment program enables policy makers to make programming decisions at the provincial, district or school level. (New Brunswick Department of Education and Early Childhood Development, 2015)
>
> [Newfoundland and Labrador:] Information obtained from provincial assessments is used to inform decisions on teaching and learning. Teachers, school administrators, school districts, and the provincial government use this information to improve student achievement. The information is also used by schools to chart student performance over time. (Newfoundland and Labrador Department of Education and Early Childhood Development, 2015)
>
> [Nova Scotia:] The objectives of Evaluation Services are to: … develop and administer student assessments to assist students to achieve outcomes; provide student achievement information to government for education decision making; help teachers understand assessment principles and practices; support school accreditation through collecting, analysing and reporting results of questionnaires, which helps to improve education decision making… (Nova Scotia Education and Early Childhood Development, 2015)
>
> [Ontario:] EQAO is an independent agency that administers assessments to measure Ontario students' achievement in reading, writing and math. The objective and reliable results from EQAO's tests complement the information obtained from classroom and other assessments to provide students, parents, teachers and administrators with a clear and comprehensive picture of student achievement and a basis for targeted improvement planning at the individual, school, school board and provincial levels. (Ontario Ministry of Education, 2015)
>
> [Prince Edward Island:] Provincial assessments are conducted yearly and tell us how well students are doing at key stages of learning… They tell us how well students are learning the curriculum, where students may need help, and how resources may be directed to improve our education system. (Prince Edward Island Department of Education and Early Childhood Development, 2015)
>
> [Québec:] The intention of this examination is to provide opportunity for students to demonstrate knowledge and competency, as well as to provide teachers the opportunity to judge literacy development. Data obtained from student performances on the tasks prescribed in this examination, in conjunction with data collected from performances during the cycle, will help the teacher form judgments about the levels of competency attained by the end of the cycle for the end-of-cycle report. (Québec Ministère de l'Éducation, du Loisir et du Sport, 2011)
>
> [Saskatchewan:] A new assessment approach will provide fair, reliable, valid and timely assessment results to students, parents, teachers, school divisions and the province, in order to monitor progress and implement plans to improve student achievement. Assessment information must be used to monitor and support student improvement, provide targeted supports to teachers, and improve teaching, learning and curriculum for all students. (Saskatchewan Ministry of Education, 2015)

The choices made at the ministry level regarding program-delivery and curriculum are policy-

level choices (see Volante, 2013), but they are not the main focus of this study. It should be clear that the ministries also have explicit and high expectations for the classroom-level use of LSA data by individual teachers.

Table 3
Reactivity Effects Shown for Each Province and Nationally

	AB	BC	MB	NB	NL	NS	ON	PEI	QC	SK
Gr. 1										EF,M [5]
Gr. 2				EF						EF,M [5]
Gr. 3	EF,M		EF,M	M	EF,M	EF	EF,M	EF,M		EF,M [5]
Gr. 4		EF,M		EF		M			F	EF [7]
Gr. 5				EF,M						M [7]
Gr. 6	EF,M,S,SS				EF,M	EF,M	EF,M	EF,M	EF,M	
Gr. 7		EF,M	M	EF						EF [7]
Gr. 8			EF	M		EF,M				M [7]
Gr. 9	EF,M,S,SS			EF [1]	EF,M		M	EF,M		
Gr. 10						EF,M	EF [1]		M,SS,S [1]	
Gr. 11				EF [4]			EF [4]		EF	EF [7]
Gr. 12	EF,M,S,SS [3,6]	EF,O [3]	EF,M [3]	EF [3]	EF,M,S,SS [3]					EF,S,M [2]

1 - Graduation requirement (must be passed)
2 - Graduation requirement only when teacher not accredited
3 - Mark on exam assigned a designated value of final grade
4 - Re-write of graduation requirement exam
5 - Piloting in the 2013-2014 school year
6 - Student must write both EF and SS
7 - Suspended since the 2012-2013 school year

EF - Core English or French
M - Mathematics
S - Science
SS - Social Studies
O - Other

Research Questions

This paper is a program evaluation study intended to examine how closely provincial assessment policy goals align with the practices of classroom teachers. As a result, the research questions posed here include:

1. How reactive are Canadian teachers to provincial large-scale assessments? Secondarily, what "type" of reactivity is most common (i.e., teaching [to] the curriculum or teaching to the test)?
2. How do teachers perceive large-scale assessment policies in the context of educational accountability? And how do they understand their professional role in this process?

In answering these questions, it should be somewhat clarified how closely aligned teachers' classroom practices are to the stated policy goals of the various ministries of education across Canada.

Methods

This paper will use mixed methods drawing upon both survey (quantitative) and interview (qualitative) data to inform the discussion. Mixed methods are often used in social sciences research in order to provide the best of what quantitative and qualitative methods can offer. Put another way, the data sets may be complementary even if they are not in complete agreement (Onwuegbuzie & Leech, 2005).

It is by no means certain or expected that qualitative and quantitative results will align. Brannen (2005) gives four possible outcomes from mixed methods including: (a) corroboration, which has the two sets of results aligned; (b) elaboration, when more information or clarity is provided; (c) complementarity, despite differing results insights are gained; and (d) contradiction, when findings conflict. Given that LSAs are controversial in the teaching community as much as in academic research, elements of elaboration and complementarity are realistic aims.

Surveys

A research survey was written and distributed using Survey Monkey that tabulated responses directly into a spreadsheet (reducing data entry errors; Nardi, 2003). Surveys were sent to teachers in all Canadian provinces and to different schools with grade levels K through 12. Field-proven questions were

adapted from various important research studies: (a) reactivity practice questions were adapted from both Skwarchuk (2004), and Hamilton and Berends (2006); (b) attitude questions were adapted from Brown (2004); (c) appropriate uses of LSA data questions were adapted from Wayman, Cho, Jimerson, and Spikes (2012); and (d) supports and professional development questions were strongly influenced by Boyle, Lamprianou, and Boyle (2005).

Surveys were sent in the 2013-2014 school year all across Canada for a cross-sectional dataset. The sampling unit was individual teachers, Canadian public school teachers were the target population, and clusters for the probability sampling were school divisions (also known as boards or districts). As a result of the clustering, the target population was already divided into non-overlapping groups (school divisions). In order to allow the two groups to be compared, all teachers from participating schools were asked to take part whether they administered LSAs or not. The selection process was affected by the reality that many school divisions chose not to participate. Cluster choices ended up as much a voluntary sample as a random one when division- and/or school-level administrators were deciding whether or not to allow distribution of the survey.

The large national dataset ensured that all strata of teachers were represented. Statistics Canada collects data on only teacher sex, and age. A comparison of the collected sample to these data showed high levels of congruity. National sample age data are 92.6% congruent with these Statistics Canada (2007) data, and the sample sex the data are 99.4% congruent. These strong congruity supports the generalizability of these findings.

The analyses are built around a conception of the dependent variable (Y) in this study:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i \qquad i = 1, , , , , n \qquad (1)$$

- Y_i is the dependent variable or reactivity teacher i (use of test data).
- X_{1i} is the first independent variable, X_{2i} is the second, etc. these are explanatory variables for the teacher i.
- intercept β_0 is the expected value of Y when all Xs equal 0.
- β_1 is the regression coefficient of X_1; β_2 is the regression coefficient of X_2, . . ., β_k.
- e_i is the residual of the regression.

For data analysis purposes binary values were applied to different responses. Respondents were asked to rate the frequency of their use of different reactivity strategies from the choices *never*, *sometimes*, or *always*. Teaching (to) the curriculum responses were respectively given values of 0, 0.5, or 1. Teaching to the test responses were scored in negative values. Respondents could therefore score between 0 and ± 5 on each scale. A test on internal consistency, Cronbach's alpha, was used for the dependent variable. The alpha score for reactivity items was 0.742.

Interviews

The interviews were done using a semi-structured format, and subjects were selected purposively in order to represent the reactivity strategies reported by survey respondents (Flick, 2009). The interview guide followed the same themes as the survey using established coding frames from surveys (as in Blaikie, 2000) in an effort to triangulate qualitative and quantitative data (Jick, 1979). Since interviews were done to explore in more depth the lines of inquiry from the survey, commonalities and details emerged to support the quantitative results. It is also true that contradictions and differences in opinion were made apparent, these indicating more accurately the wide range of perspectives on the topics covered. The number of interviews was relatively small but included teachers from across Canada and several school- and division-level administrators to determine their level of engagement in LSA reactivity. Subjects were assured anonymity and confidentiality but are identified by subject and grade level taught, province, and sex.

Interviews were conducted with 13 classroom teachers (all of whom administer LSA tests), 10 in-school administrators, and four division level staff. Only classroom teachers would have previously completed the survey, but both in-school and division level administration were also included to gauge the alignment of responses from front-line staff with their supervision teams. As with survey respondents, these subjects were assured anonymity and confidentiality.

Limitations

The surveys had a reasonable response rate for email-based data collection, yet it might be argued that

non-response bias could be present. The sample of respondents covered all ten provinces in Canada, but only a small n for each province was garnered. This makes any in-depth province-level analysis problematic, but some general provincial findings are included. The sample also was subject to the permission needed from school divisions to conduct the research study. As a result, disparity between urban and rural participation became noticeable in some provinces (for example in Saskatchewan neither of two large city centres were willing to participate). The sample of interviews was also quite small, but being chosen purposively to explain the relationship between reactivity effects to accountability policy, the researcher still asserts that these data have value to clarify and augment the quantitative results.

Findings (1) – Quantitative Survey Data

Teachers are Reactive to LSA Data

Response rates to the emailed survey were at statistically acceptable levels, and considering the number of surveys that are presented in a given year to teachers, quite respectable. Participation entailed: (a) 27 school divisions; (b) 181 schools; (c) 5523.1 teachers (based on FTE); (d) 1071 responses; (e) 453 responses from test-administering teachers; and (f) an overall 19.4% response rate (42.3% of these respondents administering LSAs). The data of primary importance for this study were from these teachers who give LSAs in their classrooms, and for the analysis, partially-completed surveys were removed from these respondent groups. Teachers self-reported the use of different instructional strategies adopted as a response to the data they may receive from large-scale assessments done in their classrooms.

Table 4
National Data for Reactivity

	Prov.	Avg. score	SD	Prov.	Avg. score	SD	Prov.	Avg. score	SD		
Rank	Teaching (to) the curriculum			Teaching to the test			Net Reactivity			*n*	
1st	AB	2.99	1.08	ON	-3.73	1.17	NS	0.29	1.40	AB	46
2nd	NB	2.92	1.05	QC	-3.54	1.11	PEI	-0.07	1.14	BC	39
3rd	QC	2.84	1.16	NL	-3.53	0.80	SK	-0.26	1.17	MB	37
4th	PEI	2.82	1.08	AB	-3.53	1.15	NB	-0.32	1.20	NB	55
5th	NL	2.72	0.90	NB	-3.28	1.03	MB	-0.38	1.38	NL	29
6th	NS	2.52	1.11	BC	-3.13	1.21	AB	-0.49	1.36	NS	48
7th	MB	2.37	1.17	PEI	-2.90	1.03	QC	-0.70	1.37	ON	51
8th	ON	2.26	1.25	MB	-2.74	1.28	NL	-0.81	1.28	PEI	34
9th	SK	2.21	1.17	SK	-2.43	1.17	BC	-1.31	1.16	QC	29
10th	BC	1.83	0.95	NS	-2.21	1.31	ON	-1.46	1.30	SK	50
	CANADA	2.50	1.15	CANADA	-3.09	1.24	CANADA	-0.56	1.37	CANADA	418

Collated in table 4, what is immediately apparent is the fact that teachers are quite reactive to these data. Teaching (to) the curriculum had a range of scores from 0 through +5, and teaching to the test was scored 0 through -5. When both teaching (to) the curriculum and teaching to the test strategies are considered we find the average national score is 5.63, which indicates that both teaching to the test and to the curriculum strategies are used commonly, often side-by-side. Net reactivity, a total of positive and negative applied values, shows the balance between reactivity strategies used. It tips heavily toward TTT practices over TTC ones (see table 4). Since the informed use of these data by teachers is an explicit element of all ministerial policies, this appears at first glance to be a good example of policy implementation aligning with its purpose.

The three most reactive provinces are, in order, Alberta, Québec, and Newfoundland and Labrador. The three least reactive are Nova Scotia, Saskatchewan, and British Columbia (see table 4). Teaching (to) the curriculum is most prevalent in Alberta, New Brunswick, and Prince Edward Island. Teaching to the test is most common in Ontario, Québec, and Newfoundland and Labrador.

TTT is more common than TTC. Equally significant to the amount of reactivity noted in table 4 is the type of reactivity employed. It is seen that teaching to the test scores are higher than those for reported teaching (to) the curriculum effects. This indicates that teaching to the test strategies are more commonly used with LSA data. The distributions in figure 1 (especially the large right tail) are also indicative of higher amounts of teaching to the test being employed. In all provinces, save Nova Scotia, TTT is more common than TTC.

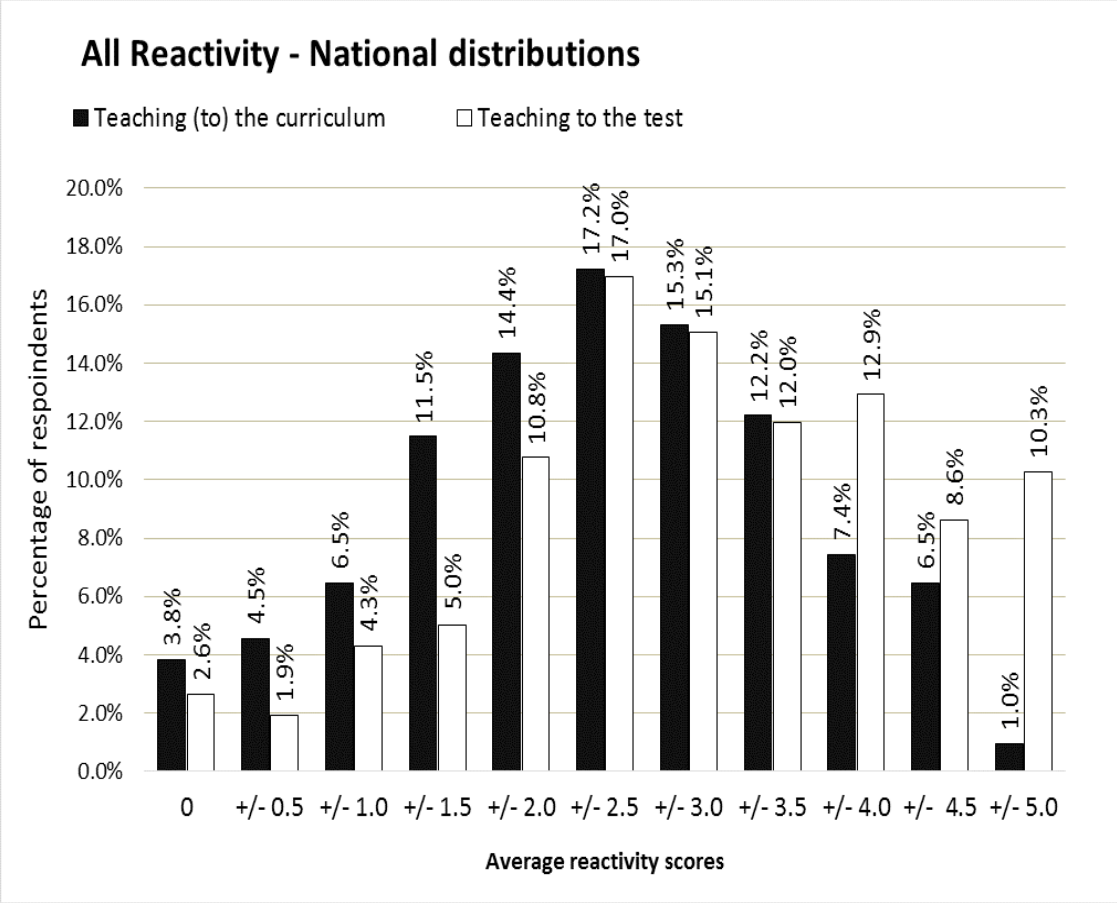

Figure 1. Subjects and grade levels assessed in Canadian provinces

A larger study of Ontario teachers' practices would be needed to make more definitive and specific observations about reactivity here, but the relatively small sample of 46 teachers clearly favour TTT over TTC. This is also true of almost all other provinces in Canada, although all to a lesser degree.

High Stakes and Minimum Competency Testing

The high-stakes dimension of LSA testing has been covered in another paper (Copp, 2018), but it does bear some mention here. Those provinces that have the highest rates of teaching to the test also give either high stakes or minimum competency tests in high school (see table 4).

High stakes tests have been shown in the literature to have a strong influence on the use of TTT strategies by teachers (Darling-Hammond & Rustique-Forrester, 2005; Mintrop & Sunderman, 2009; Volante & Cherubini, 2007). Higher total reactivity from provinces that have such tests is thought to be an indication that reactivity in general and TTT in particular is most common where high stakes testing is practiced.

The three provinces that do not currently have high stakes or minimum competency exams at the high school level are Prince Edward Island, Saskatchewan, and Nova Scotia. For these jurisdictions it is clearly relevant that they also have the most comparable scores between reactivity types. There is a near balance between reactivity scores for these provinces, although that is perhaps not reason enough to celebrate.

Lower total reactivity scores might be an indication that more defensible types of reactivity are employed and less TTT. We should not ignore the fact that it absolutely does matter what kind of instructional change assessment policy is promoting since not all change is in the best interests of well-rounded learning or of a supportive educational environment.

Findings (2) – Qualitative Interview Data

With the reactivity results shown thus far, it is clear that accountability policies do not effectively guide teachers toward TTC strategies. Teaching to the test is more common in nine of 10 provinces and clearly the more popular choice in the national sample. Taking provincial LSA policies at face value, this cannot be seen as anything other than a failure of implementation since narrowly focused test preparation is hardly the same as improved teaching. Hearing directly from reactive teachers will allow us to see this more clearly and make it possible to recommend actions for moving forward. We can take direction from Duemer and Mendez-Morse (2002):

> The qualitative scholar can investigate informal communication lines to develop an understanding of how policy is developed, implemented, and how it changes in the interim. A people-centered focus encourages us to better understand the role of individuals throughout the institutional hierarchy in implementing policy and the influence they have in determining its final form. (p. 7)

The process involved coding to break down interview data and reformulating them in a new way. In the case of the open interviews from this study, the interviews followed themes devised to gather information about different policy nudges (supports or incentives, for example) employed to guide teachers toward the use of LSA data to improve their classroom instruction. From these pre-established coding themes, axial coding was done to identify material most relevant to the themes of accountability and reactivity (Flick, 2009). Based upon the research questions, two main themes were identified from interviews that relate to: (a) the accountability function of LSA tests, and (b) how teachers change their instruction in order to meet what they perceive to be the assessment policy goals.

Accountability Theme (1) – Policy-Level

Responses from interview subjects related to accountability, one of the key political goals of LSA testing, can be broken into two areas of concern. Where this could be seen to directly affect teachers' work was in terms of school-based accountability considerations. The less direct influences, the more distant workings of the policy are termed policy-level considerations (table 1 has LSA policy objectives delineated in the same manner).

Interview subjects had differing opinions on policy-level educational accountability, especially considering the means by which it is currently measured (using LSA results). Criticisms tended to focus on the idea that LSA was politically motivated and not really geared towards improving education. Still, not all respondent opinions were negative, and the idea of accountability rang true for some.

> The problem with that is, of course, is that [provincial assessment] presupposes or has for a basic premise, is that a significant number of teachers in this province teaching senior subjects don't do valid evaluation in their classrooms and that a three hour exam is a better judge of a student's achievement than my 10 months of evaluation. (QC, High school English teacher, male)

> In all honesty, I think we're public employees, and I prefer transparency overall… I think the majority of the people I have worked with say, "You know what, whatever goes on in my classroom ought to withstand public scrutiny." (AB, High school Science teacher, male)

The kind of scrutiny that was widely criticised was the publication of school rankings, (especially when done by third-party think tanks such as the Fraser Institute). When results are poor, they often come along with public or media cries to revamp curricula, the assessments themselves, or provide more resources or staffing to schools.

> That's it, the Fraser Institute … published results in the newspaper, and says these are the top schools and these schools aren't doing very well. Our particular school, we are in an area with

> high socio-economic needs. And there is another elementary school in my community and cater, like a lot of the doctors and research scientists who live in that area, so there their school usually does really well. (BC, Elementary homeroom teacher, female)

> The only piece I would struggle with is if it was something that was publicized, you know, put in a local paper… and then people might say, "Oh, what's wrong with [school name], and who is the Math teacher there?" You know, laying blame on staff. (PEI, High school Math teacher, male)

Another commonly cited purpose of LSAs was to confirm curricula are being taught in classrooms. Mixed feelings about this oversight were evident.

> I do think, [that] for all the wrong reasons, that provincial exams at the grade 10, 11, 12 level create a sharper focus for the teachers to pay more attention to those courses and because they know those results are going to be higher stakes and be made public… I think the aspect of accountability, combined with the impact on kids and potentially their futures really does improve the teaching in those courses. (BC, Division staff, male)

> I think that provincial assessments, umm, force is not the right word, but, direct, umm, teachers more to the curriculum documents to and to follow the curriculum documents closer than maybe they have in the past, yeah. I guess maybe it is an accountability kind of a thing, I don't know, but we're expected to, you know, to use the curriculum documents and to ensure that all the material is covered in the curriculum, so. Without provincial assessments I am not quite sure it would be that focused. (PEI, K-9 school vice principal, female)

Accountability Theme (2) – Classroom-Level.

It was not always the case that interview respondents could see the link between provincial accountability and their classroom practices. These classroom-level considerations have a more direct influence on practice than the ideological opinions about policy-level accountability. Many teachers and administrators agreed that there was an element of uncertainty about the link between these large-scale mandatory assessments and practical changes to classroom instruction:

> No, I don't think that the average teacher sees a link between provincial testing and their classroom instruction… Classroom instruction is the number one predictor of student achievement and that's what needs to happen. No, I don't see any link between the results on the EQAO test and teachers' classroom instruction. I don't at all in any subject, division, grade, not at all. (ON, High school English consultant, female)

> If the department of education is saying, well, we need to do these assessments so at the department level they can, it can help them direct their curriculum and the direction going forward, that is one thing… But I just don't feel that it is as beneficial here within the building… Just getting the assessment results and putting them on the teacher's desk and continuing along isn't necessarily a help to the teacher. (PEI, K-9 school principal, male)

> Many educators were concerned that the assessment instrument (or the way it is administered) was not designed for classroom use or adjustments to teaching practices at this level…. It is unfortunate that kids have the test in February when I still have March, April, May and June to cover the curriculum. So there are questions on the test that I haven't even covered… So that part is frustrating as well, that you can't cover all of the curriculum by February and have them assessed. (BC, Elementary homeroom teacher, female)

> There are some outcomes that lend themselves to more easily being assessed in a machine-scored way. So I try to make sure that I focus on having the best, most recent information on what my kids know especially about those things that cannot be assessed on that test. So that their blended mark is a balance of those two things. So that their classroom mark is

> not a predictor of their diploma exam mark but instead an assessment of the other things. (AB, High school Math teacher, female)

Some educators were inclined to appreciate the "external check of standards" function that LSAs serve. It was seen as a "level-playing field" and a check on the grade inflation that had increasingly tainted high school marks.

> Probably the best feature of it is that it is kind of a level playing ground for the students. We have a fair bit of grade inflation in some of the schools in the province based on, I think it is parental pressure but also the pressure of getting into post-secondary. So what the exam does, it kind of gives an equal chance for all the students… Some teachers mark easier than others, some teachers mark differently, so this is one of those ones that provides a fairly uniform testing process. (AB, High school Science teacher, male)

> "Certainly, from an accountability perspective, it certainly gives you a sense of how your kids are learning in your building compared to kids in other buildings, not just in your jurisdiction, but, you know, it compares apples to apples" (ON, High school principal, male).

It can be fairly said that there is a range of opinion about the classroom-level accountability function of LSAs across Canada. There are various identified benefits, but the distribution of these between provinces, even between school divisions is highly irregular. The question (not addressed directly in this study) of whether a single assessment tool is capable of providing quality information at the policy- and classroom-levels is the elephant in the room for this discussion.

Reactivity Theme (1) – Teaching (to) the Curriculum

Comments on teacher reactivity can also be divided into two main sub-topics. Reactivity effects can be seen from the context of TTT or TTC. The survey prompts themselves seemed an appropriate way to organize and code reactivity-based responses.

Starting with TTC strategies, many teachers had the expectation of test-based Professional Development (PD) as a means to guide their instructional change (reactivity). This was available in some locations, but certainly not all. It was also true that some teachers chose not to take advantage of the offered PD.

> That is the piece that is missing is that, Okay so now we have done an assessment, now we see what has happened, or we don't see what has happened very well. Now what? What does it all mean? Teachers are like students - if they don't see purpose in what they are teaching or doing, it's not productive. (MB, Elementary school principal, female)

> There wasn't PD on it. There wasn't anything. I don't think so, there wasn't any PD or anything. We were just given it, it was given to us, go ahead and use it and do it, show it to your students... And a little bit of collaboration with the other grade 7 teacher to say "let's do it on this day." (SK, Middle years homeroom teacher, female)

> Not everybody does [access PD support]. You know the reality if Physics is not the only subject I am teaching, if I'm teaching Chemistry and Phys. Ed. and whatever else. . . Maybe I don't have time to take part in that, and maybe it's not my first priority. I would say the opportunity is there, it may not be taken all the time for various reasons. (AB, High school Science teacher, male)

Supports for successful implementation would also usually mean providing related resources for teachers.

> I think it depends, first of all, on how the assessment is constructed and, number two, what you do with the results of those assessments and if they impact on the actual day-to-day planning and interventions. And if you can take those and say, "Yes, it has made a difference and we were able to get the resources that we needed. We were able to do the things we wanted to do."

> Then, I think, we can say yes, they were useful. (NS, Division staff, female)

> The Math we're doing next year, it looks like it is going to be very good. But all the grade three teachers coming together, we're all getting the same message, we're all going to be given the same resources, so in theory, we can all do the same job. (PEI, Elementary homeroom teacher, female)

> In response to the wide-spread use of teaching to the test strategies, which many respondents criticized as a poor pedagogical practice, some teachers consciously chose a different path…. I'm going to try and say this very carefully. We care very much about how our students do; we very much want our students to succeed. I think we feel that the provincial exam and how they are dealt with are almost completely out of our control… So, we teach what we know the students need to know, try to have in as much of the other things that we think they should know to be well-rounded people and then we kind of let the chips fall where they may when the exam comes around. (QC, High school English teacher, male)

> Some of my colleagues are more interested in improving student understanding and the belief there being that if our students understand better then the test format and wording and strangeness of questions and such will not be as big an issue because they will know the content. And other teachers focus on, we need to ask questions in this way… so students aren't tricked by asking this question in this way. We need to teach kids how to write this type of test, which I think are "improving scores" conversations. So it depends very much on the specific teacher and their particular philosophy. (AB, High school Math teacher, female)

> Making sense of the LSA results data, itself no simple task, was commonly reported to be a job best done with teaching colleagues. Without this support, interpretation and implementation may not be done to high standards. …I think by collegial development and sharing and all of that, you become a better educator. There is lots of PD out there, but there is no greater PD than working with your colleagues. (NB, High school principal, male)

> This is my third year in this particular situation, preparing students for a grade 9 assessment. And each year they have had three to four meetings throughout the school year where you would sit down and discuss, you know … the topics covered, the strategies in particular areas based on what the province would see where students might have struggled in past years. (PEI, High school Math teacher, male)

> In the past I've been at larger high schools so I would have colleagues who taught Physics as well so we would look at our common outcomes and say maybe here is something… You know it would lead to actual professional development. (AB, High school Science teacher, male)

Reactivity Theme (2) – Teaching to the Test

Of the five teaching (to) the curriculum strategies covered in the survey (see table 2), only four were all well represented in interviews. Respondents did not mention the last strategy, holding extra study sessions after school, and the researcher did not employ a leading prompt.

Looking now at the teaching to the test strategies, respondents were very honest in admitting their preferential use of these practices, and in some cases made it very clear that this was the expectation at their school.

> I will be really honest with you. I have been kind of responsible for overseeing [OSSLT] preparation… I treated it as a very custodial exercise. I mean, I looked at the test and I sort of, ahh, I put together lesson plans that I thought any teacher would be able to deliver to their students… we do all of our training in grade 10 classes period 1, because those are the kids you have to hit and I want teachers to have ownership. So what happens is we teach to the test.

(ON, High school principal, male)

> What's the expression about the tail wagging the dog? The diploma exam has so much pressure on it that people are teaching in ways that they know are pedagogically unsound because it will get kids through the test… I have sat in staff meetings where people will justify, "I know it's not the best, but it gets them through the test." So absolutely, I would say it has a detrimental effect on classroom teaching. (AB, High school Math teacher, female)

Respondents witnessed or participated in curriculum narrowing because the year-end test had become, in some cases, the *de facto* curriculum. "You have to make sure that you have covered most everything. I would say the things that comes back the most are fractions, geometry… basic facts" (QC, Middle years homeroom teacher, female). "If there were no assessment, I wouldn't have my students pretending they are writing for newspapers and trying to write scripts and those kind of things" (NB, Middle years homeroom teacher, female).

> I know there are teachers who teach to the test and cover graphing to make sure that the kids know about graphing because the graph question is worth four points. But I don't believe in teaching to the test. I like to cover the material as I think my kids are ready for it. (BC, Elementary homeroom teacher, female)

> I know the previous year to that we were really bad on quadratic equations. . .We didn't put enough emphasis on, I know I didn't - I may have done three examples. So this year I nailed a quadratic equation, and we solved that problem. (NL, High school Science teacher, male)

One particular form of curriculum narrowing is teaching test-taking strategies. There are no curriculum outcomes in Canada at any grade known to the researcher that spell out the process of elimination or answering on machine-scored cards as objectives.

> [My students] learned a lot more than they would have if we had had the PATs [Alberta LSAs] this year because I could actually teach the curriculum and the students instead of trying to worry about teaching them how to fill in Scan-tron bubbles and how to write a copy-cat story. (AB, Elementary English teacher, female)

> We took the time to say, okay, these are the types of questions you are going to get the types of what you will actually see, not content-wise. So, I don't look at that as teaching to the test, but you have to teach to the style of the test… So we took a lot of time to look at what, how are they going to ask the questions and, ahh, and what types of response will they be looking for. Now you could say ... that is teaching to the test but it isn't because we have no idea what the content is like. (NB, High school principal, male)

It was also common practice for teachers who administer LSAs in their classrooms to alter their own assessment practices to mirror the format of the test. No doubt there were fewer surprises on the test that way, but it also limits the creativity of the teacher (and the student).

> It is an inappropriate outcome that comes from the exam but when the teachers choose to do it from the best intentions, to help kids out, I can't say it is inappropriate. I can say it is a negative consequence of having a provincial exam that it is making people feel they need to do that. (BC, Division staff, male)

> Most of our curriculum is getting kids to make connections with the text and we are not encouraged to get the kids to read a book and answer questions. But when it's time to do an assessment, they have to read a book and answer questions. But that is a learned skill, you have to take time to do that. (PEI, Elementary homeroom teacher, female)

> We have three weeks of every single day going to teach them how to … manage multiple choice tests and every single item in the acids and bases is to be covered. And then after seeing

> five or six of the same types of questions, they're ready for it. And then they'll know exactly what they are expecting. So, their marks are high, everybody's happy, they're getting their scholarships, they get their entrance marks. (NL, High school Science teacher, male)

The content of tested subjects tends to take precedence over other things in classrooms that have LSAs. In-subject narrowing occurs, as well as missing out on non-core material.

> … just because you have that thing at the end of the year that you are kind of, "Oh my goodness! I have to make sure my kids are ready for this." But it detracts in some ways because you feel like, well I have to cover x, y and z, so I don't have time to do a class play this year… Not that covering curriculum is a bad thing but if you got really took up in it you wouldn't give the kids any of the other experiences that also round out their education. (PEI, Elementary homeroom teacher, female)

> I think we let go a lot of conventions like spelling, I think. Not me, but I think teachers in general don't do [vocabulary] as much, and maybe not do spelling and grammar as much because there is no time. Because we have to work so much on response because at the end of the year, that is what it is about… But I think it is so important for them to learn to write properly. The exam doesn't leave any room for that. You don't correct spelling during those exams, you don't correct it. (QC, Middle years homeroom teacher, female)

The final teaching to the test prompt related to using old exams as a means of preparing for the current one. This was a practice that was reported to be commonly used.

> I think the worst that I see … is when people simply use past diploma exams to study by rote. I think it is rather ineffective. There is very little crossover from one exam to another… That would be the, not necessarily abuse, but the less effective use of diploma exams. (AB, High school Science teacher, male)

> I'm a proponent of it, and you may say that I teach to the test, which I definitely do… We are rather an elite school and they're concerned about scholarships, they're concerned about entrance marks and they want to do well when they go to university so it all centers around this test. By doing the test, by teaching to the test, by teaching every single item… like I said I've got every single test up there on my website and I go through them all. (NL, High school Science teacher, male)

> I don't think it is appropriate for a teachers to get old FSA exams [British Columbia LSAs] and teach to that… Whereas when it starts counting, if you will, towards the kids' marks and their future and you know that this is a reality that the kids are facing I would say that it is appropriate, not necessarily the best educational thing ever, but it is appropriate because teachers are supposed to help kids. (BC, Division staff, male)

These qualitative data go a long way to explaining some of the different motivations and rationales teachers have for using data in the manner they choose to do so. Especially as it relates to high-stakes exit exams, it is difficult to condemn teaching to the test practices when teachers know that scholarships, university entrance, and other important academic rewards are based on these heavily weighted provincial tests. As long as LSA policy paints teachers into this corner, you will find teaching to the test. It is also the case that as long as reflective, professional-minded educators are in classrooms, they will do everything they can to ensure the success of their students. What that means, in practice, is still a matter for debate and an opportunity for future research study.

Conclusions

Program evaluation studies are designed to confirm that a given policy or program has been successful based on the goals defined within the policy itself. This paper has done nothing more than ask if teachers are using the data from provincial LSAs as policy dictates, and then if the manner in which is being used is in accordance with widely accepted best practices for both the evaluation and the education of students.

The findings have shed some light on assessment practices and important lessons can and should be taken from the study.

Reactivity is Present

The first and most important point to make is that teachers are acting on the data they receive from provincial tests all across Canada as are the expectations of ministerial policies. Average reactivity scores are significantly high values.

These results show that provincial assessment policy has been successful in influencing teachers' instruction using LSA data. There is more to successful implementation, though, and that will be addressed next.

Types of Reactivity

In terms of the types of reactivity employed, not all changes in instruction are necessarily for the better. Instructional change can improve the education system, but only if the movement is consistently in a direction that aligns with accepted pedagogy and ministerial goals. Teaching to the test strategies are an effective means to improve test scores, but little else is accomplished by their wide-spread use. Teaching (to) the curriculum strategies have more leverage across learning domains, focus on more outcomes, and allow students a greater variety of ways to demonstrate their learning.

Assessment policy-makers should know that teachers will change their practices if that is the expectation, but the specific nature of the expected changes needs to be made more explicit and must be implemented more uniformly across school divisions and schools. This means filling the gaps in professional knowledge with PD, resources, and support.

Accountability

The connection between LSAs and classroom teaching is not completely clear to many teachers. They understand and may even support the policy-level generation of data for evaluation of the education system, but they often do not think this same data set is well-suited to the improvement of classroom teaching. Respondents were almost uniformly opposed to the trial of teachers and schools in the court of public opinion, as wide-spread publication and third-party evaluation tend to inspire.

Keeping the education system accountable for itself seems not to sit comfortably alongside having teachers improve their instruction. Reactivity is a change in behaviour resulting from external evaluation, and teachers are reactive. The difficulty lies in fostering TTC strategies over TTT strategies. There are other ways to achieve accountability goals without high-stakes or even census-style tests. As an example, low stakes tests (for improved school-level information), and sample-style tests (for policy-level information) are widely used outside Canada.

Reactivity

Looking at the quantitative data on reactivity types and prevalence, it is somewhat tempting to judge teachers as acting improperly or of being poorly trained. The interview respondents provided details and rationales for their actions that made such facile conclusions hard to support. There is a wide rift in opinion about the appropriateness of teaching to the test, but it is clear that provincial LSA policies make explicit the directive for teachers to use the data despite the fact that provinces, according to many respondents, have not provided enough of the resources, supports, and individualized direction needed to reach policy goals.

Advocates of the continued and expanded use of large-scale assessment aspire to having tests that are "worth teaching to" but regardless of the quality of the instrument used, assessment policies need to address the fact that many evaluation data are not being used as it was thought they would be. Policy levers have proven to be quite effective at inspiring teachers to change their instruction, but it is not clear to many teachers in what direction their instructional "course-corrections" should be made. At the most basic level, practice has to follow belief, and the level of confidence teachers have in the assessment tools they have been provided by the provinces is not sufficient to support the belief that instructional change can and should be guided by the data from these tools. More groundwork is needed to show educators that

high-quality, curriculum-aligned LSA has the potential to improve their teaching practices.

References

Alberta Education. (2015). *Achievement testing results.* Retrieved from https://education.alberta.ca/provincial-achievement-tests/about-the-pats/, 1

Ashton, P., Buhr, D., & Crocker, L. (1984). Teachers' sense of self-efficacy: A self-referenced or norm-referenced construct? *Florida Journal of Educational Research, 26*(1), 29-41.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267. doi:10.3102/0013189X07306523

Benveniste, L. (2002). The political structuration of assessment: Negotiating state power and legitimacy. *Comparative Education Review, 46*(1), 89–118.

Blaikie, N. (2000). *Designing social research.* Cambridge, UK; Malden, MA: Polity Press.

Bolon, C. (2000). School-based standard testing. *Education Policy Analysis Archives, 8*(23), 1–43.

Boyle, B., Lamprianou, I., & Boyle, T. (2005). A longitudinal study of teacher change: What makes professional development effective? Report of the second year of the study. *School Effectiveness and School Improvement, 16*(1), 1-27.

Brannen, J. (2005). Mixing methods: The entry of qualitative and quantitative approaches into the research process. *International Journal of Social Research Methodology, 8*(3), 173–184. doi:10.1080/13645570500154642

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers, 71*. doi: 10.1787/5k9fdfqffr28-en

British Columbia Ministry of Education. (2015). *Transforming assessment*. Retrieved from https://curriculum.gov.bc.ca/assessment

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301–318. doi: /10.1080/0969594042000304609

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297.

Copp, D. T. (2015). *Teacher-based reactivity to provincial large-scale assessment in Canada.* Boekenplan: Maastricht.

Copp, D. T. (2016a). Teaching to the test: A mixed methods study of instructional change from large-scale testing in Canadian schools. *Assessment in Education: Principles, Policy & Practice, 1*(20). doi:10.1080/0969594X.2016.1244042

Copp, D. T. (2016b). The impact of teacher attitudes and beliefs about large-scale assessment on the use of provincial data for instructional change. *Education Policy Analysis Archives, 24*(109). doi:10.1080/0969594X.2016.1244042

Copp, D. T. (2018). Policy incentives in Canadian large-scale assessment: How policy levers influence teacher decisions about instructional change. *Education Policy Analysis Archives, 25*(115). Retrieved from http://dx.doi.org/10.14507/epaa.25.3299

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education, 104*(2). doi:10.1111/j.1744-7984.2005.00034.x/pdf

Duemer, L. S., & Mendez-Morse, S. (2002). Recovering policy implementation: Understanding implementation through informal communication. *Education Policy Analysis Archives, 10*(39), 1–11.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology, 113*(1), 1-40.

Flick, U. (2009). *An introduction to qualitative research* (4th ed.). London, UK: Sage Publications.

Fullan, M. (2011). *Choosing the wrong drivers for whole system reform.* East Melbourne: Centre for Strategic Education Seminar Series, 204. Retrieved from http://www.michaelfullan.ca/

home_articles/SeminarPaper204.pdf

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction.* (Consortium for Policy Research in Education (CPRE) Report #RR-65). doi:10.1037/e546712012-001

Halverson, R., & Thomas, C. N. (2007). *The roles and practices of student services staff as data-driven instructional leaders.* (Wisconsin Center for Education Research (WCER) Working Paper No. 2007-1). doi: 10.1080/00131720709335008

Hamilton, L. (2003). Chapter 2: Assessment as a policy tool. *Review of Research in Education, 27*(1), 25–68. doi:10.3102/0091732X027001025

Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments.* (RAND Education Working Paper WR-374-EDU). Retrieved from http://www.rand.org.cn/content/dam/rand/pubs/working_papers/2006/RAND_WR374.pdf

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24*(4), 602–611.

Johnson, P. E., & Chrispeels, J. H. (2010). Linking the central office and its schools for reform. *Educational Administration Quarterly, 46*(5), 738–775. doi:10.1177/0013161X10377346

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources, 37*(4), 752-777.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22*, 18–26. doi:10.1111/j.1745-3992.2003.tb00124.x

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education, 104*(2), 1-22. Retrieved from http://cse.ucla.edu/products/reports/r655.pdf

Koretz, D., & Jennings, J. L. (2010). *The misunderstanding and use of data from educational tests.* Retrieved from https://www.spencer.org/sites/default/files/pdfs/Koretz--Jennings-paper.pdf

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255.

Manitoba Education. (2015). *Assessment and Evaluation.* Retrieved from http://www.edu.gov.mb.ca/k12/assess/

Matheson, L. N., Rogers, L. C., Kaskutas, V., & Dakos, M. (2002). Reliability and reactivity of three new functional assessment measures. *Work: A Journal of Prevention, Assessment and Rehabilitation, 18*(1), 41-50.

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). Implementing data-informed decision making in schools—Teacher access, supports and use. *US Department of Education, Office of Planning, Evaluation and Policy Development.* Retrieved from http://files.eric.ed.gov/fulltext/ED504191.pdf

Mintrop, H. (2003). The limits of sanctions in low-performing schools: A study of Maryland and Kentucky schools on probation. *Education Policy Analysis Archives, 11*(3), 1–30.

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement - and why we may retain it anyway. *Educational Researcher, 38*(5), 353-364.

Møller, J. (2009). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational Change, 10*(1), 37–46. doi:10.1007/s10833-008-9078-6

Morris, A. (2011). *Student standardised testing: Current practices in OECD countries and a literature review.* (OECD Education Working Papers, 65).

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education, 25*(4), 262-279.

Nardi, P. (2003). *Doing survey research: A guide to quantitative methods.* Boston, MA: Pearson Education.

New Brunswick Department of Education and Early Childhood Development. (2015). *Provincial mathematics assessment program: Information bulletin*. Retrieved from https://www.gnb.

ca/0000/publications/.../InformationBulletinFeb2011.pdf

Newfoundland and Labrador Department of Education and Early Childhood Development. (2015). *Provincial assessments.* Retrieved from http://www.ed.gov.nl.ca/edu/k12/evaluation/crts/index.html

Nova Scotia Education and Early Childhood Development. (2015). *Nova Scotia assessments.* Retrieved from http://plans.ednet.ns.ca/about-plans

Ontario Ministry of Education. (2015). *Education quality and accountability office (EQAO).* Retrieved from http://www.eqao.com/AboutEQAO/AboutEQAO.aspx?Lang=E

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375–387. doi:10.1080/13645570500402447

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

Prince Edward Island Department of Education and Early Childhood Development. (2015). *Provincial Assessment Program.* Retrieved from http://www.gov.pe.ca/eecd/studentassessment

Québec Ministère de l'Éducation, du Loisir et du Sport. (2011). *Information document, mandatory examination: English language arts.* Retrieved from http://www.learnquebec.ca/export/sites/learn/en/content/curriculum/languages/ela/documents/InfoDoc_ELA_Cycle3Prim.pdf

Saskatchewan Ministry of Education. (2012). *Student achievement initiative: Background information.* Retrieved from http://www.education.gov.sk.ca/student-achievement-annoucement-backgrounder

Saskatchewan Ministry of Education. (2015). *Help me tell my story.* Retrieved from https://holisticassessment.gov.sk.ca/about-the-assessment/

Saskatchewan Teachers' Federation. (2015). *Code of professional competence.* Retrieved from https://www.stf.sk.ca/portal.jsp?Sy3uQUnbK9L2RmSZs02CjVy0w7ZkI/ks6g2u00gzAtsk=F#https://www.stf.sk.ca/portal.jsp?S3ua0P4leiBvLe5BSdsr0vZGZJmzTYKNX8t/KNvKOzGyZacpsswpYUA==F

Shore, C., & Wright, S. (2000). Coercive accountability: The rise of audit culture in higher education. In M. Stratham (Ed.), *Audit cultures: Anthropological studies in accountability, ethics and the academy* (pp. 57-89). London, UK: Routledge.

Skwarchuk, S-L. (2004). Teachers' attitudes toward government-mandated provincial testing in Manitoba. *The Alberta Journal of Educational Research, 50*(3), 252–282.

Statistics Canada. (2007). *Education indicators in Canada: Report of the Pan-Canadian education indicators program 2007.* Retrieved from http://publications.gc.ca/collections/collection_2007/statcan/81-582-X/81-582-XIE2007001.pdf

Ungerleider, C. (2006). Reflections on the use of large-scale student assessments for improving student success. *Canadian Journal of Education, 29*(3), 873-883.

Van Thiel, S., & Leeuw, F. L. (2002). The performance paradox in the public sector. *Public Performance & Management Review, 25*(3), 267–281.

Volante, L. (2005). Accountability, student assessment, and the need for a comprehensive approach. *International Journal for Leadership in Learning, 9*(6).

Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy, 58*, 1-21.

Volante, L, (2013). Canadian policy responses to international comparison testing. *Interchange, 44*(3-4), 169-178.

Volante, L., & Ben Jaafar, S. (2008). Educational assessment in Canada. *Assessment in Education: Principles, Policy & Practice, 15*(2), 201-210.

Volante, L., & Cherubini, L. (2007). Connecting educational leadership with multi-level assessment reform. *International Journal for Leadership in Learning, 11*(12).

Wayman, J. C., Cho, V., Jimerson, J. B., & Spikes, D. D. (2012). District-wide effects on data use in the classroom. *Education Policy Analysis Archives, 20*(25). Retrieved from http://dx.doi.org/10.14507/epaa.v20n25.2012

Wayman, J. C., & Stringfield, S. (2006). Technology-supported involvement of entire faculties in

examination of student data for instructional improvement. *American Journal of Education, 112*(4), 549-571.

Zigo, D. (2001). Constructing firebreaks against high-stakes testing. *English Education, 33(*3), 214-232.