Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
EPAA@asu.edu
Arizona State University
Estados Unidos

Copp, Derek T.
Policy Incentives in Canadian Large -Scale Assessment: How Policy Levers Influence Teacher Decisions about Instructional Change
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 25, 2017, pp. 1-36
Arizona State University
Arizona, Estados Unidos

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 25 Number 115 November 20, 2017 ISSN 1068-2341

Policy Incentives in Canadian Large-Scale Assessment: How Policy Levers Influence Teacher Decisions about Instructional Change[1]

Derek T. Copp
Good Spirit School Division
Canada

Citation: 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). http://dx.doi.org/10.14507/epaa.25.3299

Abstract: Large-scale assessment (LSA) is a tool used by education authorities for several purposes, including the promotion of teacher-based instructional change. In Canada, all 10 provinces engage in large-scale testing across several grade levels and subjects, and also have the common expectation that the results data will be used to improve instruction in classrooms. Yet despite agreement between ministries that instructional change based on LSA results is a positive development and employs data-based decision making at its heart, there remain significant differences in the kinds of incentives written into assessment policies in Canada. It is also true that implementation of the policies is less than uniform between schools and school divisions. Using mixed methods (survey data and follow-up interviews), this study examines which policy factors have the most significant impact on teacher decisions regarding the use of data. The findings indicate that highly incentivized policies correlate well to instructional change, including aspects of

[1] Due to extenuating circumstances, and with the permission of the author and the lead editor, this article was published post-peer review and revisions, but prior to a final proof by the author or article metadata translations into Spanish and Portuguese (translations may be available upon request).

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 29/7/2017
Revisions received: 9/25/2017
Accepted: 26/9/2017

both teaching (to) the curriculum as well as teaching to the test. Since the latter will be examined as a neither an educationally defensible practice nor a stated policy goal, the statement that 'incentives work' does not fully capture the nature of these impacts.
Key words: Educational assessment, decision making, reactivity, instructional improvement, education policy, Canada

Policy Incentives in Canadian Large-Scale Assessment: How Policy Levers Influence Teacher Decisions about Instructional Change

Large-scale assessment (LSA) is a tool used by education authorities for several stated policy purposes. In Canada, all 10 provinces engage in large-scale testing across several grade levels and subjects, and also share the common expectation that the results data will be used to improve instruction in classrooms. From policy-level purposes for testing (which include such goals as improving curriculum) down to classroom-level goals (like helping to identify struggling students), education ministries in Canada have been quite ambitious. Between five and nine distinct policy goals have been included in provincial education ministry literature (Copp, 2016b). Setting expectations for this many goals using a single assessment instrument is quite a formidable challenge (Schildkamp & Kuiper, 2010; Ungerleider, 2006).

In order to prod the education sector toward meaningful improvement, compliance with the mandates of LSA policies is necessary. It has long been the position of New Policy Management advocates and education economists that incentives are the most effective means to insure compliance and to achieve controlled changes in behaviour (Dee & Wyckoff, 2013; Hanushek & Rivkin, 2012; Woessman, 2001). Incentives of various kinds have been employed in this pursuit but with only mixed results (Ballou & Springer, 2015; Chambers & Tate, 2015; Corcoran, 2010). The most commonly utilized incentives for teachers and/or schools have included: a) pay or funding based on 'merit' (Koedel, 2009; Rockoff, 2003); b) earned autonomy (Morgan, 2009; Schildkamp, Poortman & Handelzalts, 2016; Wößman, 2003); c) threats of sanctions (Firestone, Mayrowetz & Fairman, 1998; Mintrop, 2003); and d) increased scrutiny or the provision of support (Møller, 2008; Stoilescu, McDougall & Egodawatte, 2016; Weinbaum, 2009).

This paper sets out to examine the effectiveness of policy incentives in the context of Canadian schools and gauge both the level and type of instructional changes teachers make in response to LSA policies. Canadian policy incentives are quite different and far less punitive than many other jurisdictions (namely the more sanction-driven models in the US and UK). Canada also does not have anywhere near the kind of private or charter school penetration seen in some other nations, so that public school policies can make a large impact. Using mixed methods (survey data and follow-up interviews) the author examined several incentives variables to determine which have the most statistically significant impact on teacher decisions regarding instruction. The findings indicate that high-stakes assessments do have a significant correlation to instructional change including aspects of both teaching to the test (TTT) as well as teaching (to) the curriculum (TTC). This correlation can be seen as a rationale for the use of policy incentives. It should be noted, however, that teaching to the test will be discussed as neither an educationally defensible practice nor a stated policy goal. The statement that 'incentives work' does not fully capture the nature of these impacts.

This paper will be laid out in the following way: a) the current literature on the use of policy incentives will be elaborated; b) the research questions will be presented; c) the theoretical model of reactivity will be explained; d) the Canadian context of this study will be explored; e) the methods

used will be examined; f) the limitations of the study will be enumerated; g) the results from qualitative and quantitative data gathering will be detailed; h) a discussion of the policy implications of these findings will be identified; and i) the paper will be summarized in the conclusion.

Literature Review

Large-scale assessment has been a fundamental tool for the purposes of educational accountability for decades in some jurisdictions and for at least 10 years across all provinces in Canada (Klinger, DeLuca & Miller, 2008). Canadian ministries have set out several specific and wide-ranging goals for these assessments (see figure 3 below). Such large-scale policies and policy goals are intended to be explicit and clear to teachers (Huber & Skedsmo, 2016; Linn, 2003). Yet having more policy goals for such assessments makes the administration of LSAs more problematic and complex. Classroom-level changes require disaggregated census-style tests while policy-level goals could use less onerous sample-style assessments much like the Pan-Canadian Assessment Program and the Programme for International Student Assessment (Hargreaves & Shirley, 2011; Morris, 2011; Volante & Ben Jafaar, 2008).

Unintended Consequences

LSA policies that promote the use of the results data must also take into consideration the instructional methods used to fulfill those expectations. Choices made about item-types and content influence teacher choices and in some cases come to supersede the full and expansive curriculum (Koretz, 2009; Luna & Turner, 2001; Shepard, 2000). The issue of curriculum narrowing is well documented and is generally considered to expend too much time and other resources on a limited scope of tested material (Holcombe, Jennings & Koretz, 2013; Volante, 2004). Limiting the manner in which students are allowed to demonstrate mastery of given outcomes is not considered the best classroom practice, but it is a key feature of standardized LSAs adopted by teachers to prepare their students (Bauer, 2000; Cizek, 2000; Datnow, Park & Kennedy-Lewis, 2012).

There are other unintended consequences of LSA. For example, when social actors are aware of the metrics that have been devised to monitor their behaviour, they react in predictable ways to have the results show them in the best possible light (Abrams, 2004; Olah, Lawrence & Riggins, 2010; Webb, 2006). This is defined in this paper as 'reactivity' which is outlined in the theoretical framework section below. In short, the more importance that is applied to a given metric, the more likely it is that reactive behaviours will be apparent. The focus of this paper is incentives, which act to increase the relative import of LSAs when pressure, stakes, or sanctions are conditional upon results scores (Altrichter & Kemethofer, 2015; Dee & Wyckoff, 2013; Hamilton & Berends, 2006).

High Stakes

Concerns about how teachers will use these data are particularly acute in cases where high-stakes testing occurs. Stakes are almost always evident to the students who must take these assessments, but this paper examines the considerable policy-related and public pressure on teachers (Spillane et al., 2002; Wößmann, 2003; Young, 2006). These professionals bear much of the blame when results are below expectations, and have justified concerns in some cases about professional consequences (Amrein & Berliner, 2002; Ben Jaafar & Earl, 2008). While not all provinces has LSAs that are high-stakes (for example only Ontario and New Brunswick have graduation requirement tests), many teachers across Canada feel the stakes involved are high as is the pressure applied to improve instruction. On the other hand, the vocal support for high-stakes testing comes from researchers who assert LSA can: ensure learning is taking place: hold schools and teachers

accountable for their work; and to level the playing field for teachers in any one jurisdiction (Allan, 2002; Bishop & Wößmann, 2004; Finnigan & Gross, 2007; Marsh, Farrell, & Bertrand, 2014; Wiliam, 2010).

Improving Instruction

All provincial ministries promote the use of LSA data to guide decision making. In general terms, the use of data to inform school-based or classroom decisions is known as data-driven decision making (DDDM). Making appropriate use of organized and meaningfully collected information, both qualitative and quantitative, helps to 'guide decisions' in schools (Schildkamp, Poortman & Handelzalts, 2016) and goes hand in glove with accountability policy. It is only reasonable to expect that data from LSAs are put to some constructive use. Datnow, Park and Kennedy-Lewis (2012) point out that being driven by data does not fix all in the education system, nor does simply accessing data change instructional practices for the better. Goertz, Oláh and Riggan (2009) support this thesis, stating that assessment literacy and professional development are necessary to make effective use of these data. An expansive perspective on DDDM is shown in Mandinach and Gummer (2016) who noted 53 skills and sources of knowledge required by teachers to effectively and appropriately use data. A summary perspective is given in Hamilton et al. (2009) in which focused on five research-approved instructional DDDM practices.

Other Testing Models

Different models of test-based accountability policies have been proposed, employed and studied. Nichols and Harris (2016) point out that lower-stakes testing (Australia), using sample-style assessments (Finland); and relying on experts via school inspections (New Zealand) are all reasonable alternatives to high-stakes, sanctions-driven policies. Altrichter and Kemethofer (2015) examined the school-inspection model in Europe and found both some positive and also some questionable reactivity effects. This was also true of Ehren and Shackleton's (2016) examination of the Dutch inspectorate. Even the Canadian and Australian 'low-stakes' models suffer from extreme public pressures after media reports of LSA data. Publication, ranking schools and the desire of teachers to appear effective tend to increase the stakes for teachers involved in testing. High-stakes leads to more well-documented unintended consequences of LSAs, namely political pressure for educational reform and the risks of teachers using gaming behaviours (Breakspear, 2012; Nichols & Harris, 2016; Rutkowski & Rutkowski, 2016).

In the end, this study sets out to focus on the provincial policy incentives which drive teacher behaviours in Canada. It is clear that policy choices can effect DDDM practices, although what drives data use may be teacher attitudes, appropriate and available support, or policy incentives. An extensive amount of research and examination of policies has been done with regards to individual schools, school divisions or provinces (Ben Jaafar & Earl, 2008; Campbell & Levin, 2009; Fullan, 2009; Hargreaves & Shirley, 2011; Levin, Glaze & Fullan, 2008; Scott, Webber, Aitken & Lupart, 2011; Volante & Cherubini, 2010; Wideman, 2002). The research study upon which this paper is based (Copp, 2015) was the first on the implications of LSA policy on a national scale ever done. This paper is one of several published and awaiting peer-review on different aspects of LSA policy.

Research Questions

This study asks how and why teachers change their instructional methods in reaction to LSA results and more specifically about the policy incentives intended to promote these changes. The research questions addressed here are:

1) Are policy incentives regarding LSA correlated to instructional change in classrooms?
2) Which incentives variables are most closely correlated with the instructional use of LSA data?
3) Are classroom-based instructional changes more strongly correlated with teaching (to) the curriculum or towards teaching to the test?

Instructional change made in reaction to LSA data is the dependent variable of this study. It had to first be established that teachers do indeed 'react' in some way to LSA results in order to move on to the more detailed questions about which incentives promote change, and what kinds of changes are made in practice.

Theoretical Framework

Reactivity explains the noticeable changes in behaviour when people know they are being externally observed or evaluated. Whether conscious or unconscious, reactivity has a direct impact on the objectivity of the metrics used. The concept was first examined in detail by Campbell (1957) as one possible design flaw in social sciences experimental methodology. In the context of this paper, it is hypothesized that teachers are reactive to LSAs at least in part because of the incentives written into related policies. The reactivity framework was chosen to provide this practical perspective on policy incentives.

Reactivity studies since Campbell's time have looked at different fields of and found the same general flaw in external assessment methods: that many people are clever enough to figure out how they are being evaluated and find ways to 'game' the results. Espeland and Sauder (2007) examined US law schools and how they were reactive to ranking tables from *US News & World Report*. Public service performance targets can also lead to gaming behaviours, as examined by Hood (2006) in the UK. Manipulation of metrics is 'the performance paradox' explained by Van Thiel and Leeuw (2002) or the 'choreographed performances' noted by Webb (2006). Whatever term is used, they are known to create distortions of reality when the performance metrics are known.

In this study, teachers were asked to respond to questions about which kinds of instructional changes were made in their classrooms in response to LSA results. They were presented 10 instructional strategies and asked to indicate the frequency of their use from the choices 'always', 'sometimes' or 'never'. A key feature of this unique model was the differentiation of these 10 strategies into two groupings based on their educational defensibility and related alignment with stated policy goals. Figure 1 shows the grouped survey prompts.

There are five strategies that closely adhere to the professional standards set out in (for example) the Saskatchewan Teachers' Federation (2015) document called their Code of Professional Conduct as well as stated ministry policy goals. These are designated as teaching (to) the curriculum (TTC). TTC includes only those practices which are considered to be both ethical and increase the number or variety of outcomes taught to students. These approaches are less likely to result in higher scores on any specific test, but they should provide the skills to improve achievement in different situations by avoiding the potential pitfalls of teaching geared to a single instrument (Popham, 2001).

Teaching to the test (TTT) includes those educational strategies which are considered to be are either unethical **or** decrease the number or variety of outcomes taught to students. TTT methods are the most direct way to improve scores on a specific test, but even those practices which might be ethical in this grouping do not have the transferability, the increased 'leverage,' upon which TTC strategies are based (Au, 2007; Cullen & Reback, 2006; Jacob, 2002; Van Theil

& Leeuw, 2002). These strategies do not meet the terms of the Saskatchewan Teachers' Federation (STF) code of conduct, especially when used to excess or by default. They also do not align well with the ministry goals as set out in policy documents (Copp, 2016b). The dependent variable of this study is the instructional change teachers make in reaction to LSA data. The use of either TTC or TTT strategies qualifies in this regard.

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 for each statement 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.

Figure 1: Reactivity prompts from the data gathering survey.[2]

The distinction made between TTC and TTT strategies was based on the judgment of the author, the terms of the STF code of conduct, as well as the education ministries' policy goals. These three considerations showed great congruence in the selection of the groups. There is a distinction made in the literature about types of reactivity, and it seemed only appropriate to treat TTC and TTT as substantially different reactions to LSA data (Koretz, 2009). It is important to note that it not being argued that the use of these strategies in specific cases is necessarily unethical. Teachers suit their instructional methods to student needs and make choices about what will work best throughout the school year, and this is just as it should be. What is suggested by the author is that the use of TTT strategies when used by default, regularity, or with entire class groups is less appropriate than using TTC strategies. The difference between the groups is one based on judgement and was not determined or expected to be shown with statistical methods. It is understood that some readers will disagree.

Canadian Context

In Canada each province has jurisdiction over education policy including which subject areas and grade levels are assessed using LSAs. Figure 2 shows the information gathered from provincial ministry websites (in 2015). It indicates that grades 3, 6, 9 and 12 have the most testing, but subjects, methods, and relative stakes for students and teachers differ. The policies also differ in terms of incentives for improvement. In each province, the results are made public either through an education ministry release and/or publishing in newspapers and other media.

[2] Adapted from Copp, 2016a, p. 7.

Policy-level choices made based on the results vary from staying the course, tweaking tests or test design, monetary support for new programs, all the way to the complete overhaul of school curricula. In many provinces, the test items are a closely guarded secret so as to avoid the risk of teachers giving students more information than they need. In these provinces teachers do not even see the tests until the day of the assessment (Prince Edward Island and Alberta are examples). Other jurisdictions freely distribute the tests and even suggest running through practice tests (Ontario and Saskatchwan are examples). There has been a furor in the press in Ontario recently about the fact that education funding has increased but provincial math assessment (EQAO) scores are going down (Maharaj, 2017). The controversy and scrutiny over scores from LSA results do have a significant policy impact.

	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)	EF - Core English or French
2 - Graduation requirement only when teacher not accredited	M - Mathematics
3 - Mark on exam assigned a designated value of final grade	S - Science
4 - Re-write of graduation requirement exam	SS - Social Studies
5 - Piloting in the 2013-2014 school year	O - Other
6 - Student must write both EF and SS	
7 - Suspended since the 2012-2013 school year	

Figure 2: Subjects and grade levels assessed in Canadian provinces.[3]
Note: Short forms for the provinces: Alberta (AB); British Colombia (BC); Manitoba (MB); New Brunswick (NB); Newfoundland and Labrador (NL); Nova Scotia (NS); Ontario (ON); Prince Edward Island (PEI; Quebec (QC) and Saskatchewan (SK).

The assessed subjects always include English or French, mathematics, and regularly science and social studies. Other subject assessments are administered only in British Columbia, and these will be discontinued in the 2016-2017 school year. There is great variation seen in the kinds of tests given to students, and what stakes are applied to teachers. Certainly there is no talk in Canada of firing teachers or principals, or of shutting schools as a result of poor LSA results. There is pressure to improve applied in each province, although the type of test somewhat dictates the kind of pressure. Public and parent scrutiny is likely the most cited by interview respondents, especially when tests are a graduation requirement or when the grades are significant factors for university admission and scholarship opportunities (more on this topic follows).

[3] Adapted from Copp, 2016b, p. 6.

Figure 3 shows the numerous policy goals set by the 10 provincial education ministries in Canada. These goals are intended to be addressed by the administration and analysis of LSAs. There is a clear tendency to include both policy-level and cassroom-level goals for these assessments.

	Purposes of large-scale provincial assessment	AB	BC	MB	NB	NL	NS	ON	PEI	QC	SK
Policy level <-------> Classroom level	Reporting/accountability	X	*	X	X	X	X	X		X	*
	Requirement for graduation	†	†	†		†	†				‡
	Reference to PISA, PCAP, PIRLS, TIMMS	X	X	X	X	X	X	X	X	X	X
	Monitoring/improving 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	Purpose evident from ministry literature
*	Not explicitly stated, but apparent from ministry literature
†	Exams must be written but need not have a passing grade
‡	Exams are mandatory when teachers are not accredited

Figure 3: Stated policy purposes of LSAs in Canadian provinces.[4]
Note: Short forms are used for the provinces: Alberta (AB); British Colombia (BC); Manitoba (MB); New Brunswick (NB); Newfoundland and Labrador (NL); Nova Scotia (NS); Ontario (ON); Prince Edward Island (PEI; Quebec (QC) and Saskatchewan (SK).

Methodology

Mixed methods are used in this study to identify both statistical correlations and also to expose explanatory details from interviews (Flick, 2006). This study employed a sequential explanatory design with a survey as the primary data collection, followed by the second phase interviews, which were analyzed in terms of the established frames from the quantitative analyses (Blaikie, 2000).

Surveys

Surveys were emailed to participating schools in all Canadian provinces and to teachers at all grade levels Pre-K through 12. A review of the literature revealed useful field-tested questions from relevant research studies including from Skwarchuk (2004), Hamilton and Berends (2006), Brown (2004), Wayman, Cho, Jimerson and Spikes (2012), Boyle, Lamprianou and Boyle (2005). Questions from these studies were adapted to meet the needs of this study. There were several themes from the literature examined: test design and data; teacher attitudes; supports for data use; and policy

[4] Adapted fromCopp, 2016b, p. 7.

incentives. Each of these is treated separately in one paper in a series from the author (Copp, 2016a, 2016b, 2017). The complete survey is shown in Appendix 2 alongside the values assigned to different responses for analysis.

All surveys were sent in the 2013-2014 school year in a cross-sectional design. The sampling unit was individual teachers, Canadian public school teachers who administer LSAs were the target population, and the probability sampling was clustered in school divisions. School divisions provided non-overlapping geographical areas of study. The selection of participants was random, yet it was subject to the reality that many school divisions chose not to take part making this as much a voluntary sample (at the division level) as a random one.

While the target population included several strata, there is a lack of demographic data available about teachers at both the national and the provincial level. Statistics Canada collects national data on only two of these strata (sex and age). National sample age data are 92.6% congruent with Statistics Canada (2007) numbers and sex data are 99.4% congruent. This level of comparability, even if for a limited number of comparators make it more likely that these results could be generalized to the wider Canadian teaching population (McMillan & Schumacher, 2010). Single province samples were too small for anything more than summary analyses, but the larger '*n*' of the nation-wide data meant that all strata were well represented.

Table 1
Response rates for the teacher survey[5]

Prov.	Number of participant divisions	Number of participant schools	Participant schools' FTEs	Teachers who may give LSAs	Responses from teachers giving LSAs	Response rate
AB	4	18	561.4	187	48	25.6%
BC	4	32	808.9	202	43	21.3%
MB	7	29	669.2	221	40	18.1%
NB	2	18	649.2	378	59	15.6%
NL	1	13	313.9	104	30	28.8%
NS	2	23	335.1	139	61	43.8%
ON	7	25	630.9	208	52	25.0%
PEI	1	28	568.4	142	35	24.6%
QC	3	13	302.6	76	30	39.5%
SK	4	40	683.7	171	55	32.2%
CANADA	27	181	5523.1	1963	453	23.1%

Note: FTEs are full-time equivalences; 1.0 is one full-time teacher. Short forms for the provinces: Alberta (AB); British Colombia (BC); Manitoba (MB); New Brunswick (NB); Newfoundland and Labrador (NL); Nova Scotia (NS); Ontario (ON); Prince Edward Island (PEI; Quebec (QC) and Saskatchewan (SK).

All teachers in a given school had the opportunity to respond (*n*=1071) but the data and analysis here come from the subset of teachers (*n*=453) who administer LSAs in their classrooms. The minimum number of respondents in this group from each province was set at 30. The overall response rate was only moderate overall (Nardi, 2006). It is uncertain, though, whether simple non-response itself casts doubt on the results of any reasonably responded-to instrument as nonresponse bias is a function of both the rate of response *and* the difference between responses for these two groups (Couper, 2000; Olsen, 2006). Response rates for the survey are given in Table 1.

The dependent variable from the larger study was teacher reactivity, of both TTT and TTC types. Other grouped variables examined were tests and test design; teacher attitudes about testing; supports to use data; and incentive to use data. Table 2 gives the average scores and standard errors

[5] Adapted from Copp, 2016a, p. 9.

from the survey reactivity questions. As a result if the low numbers of respondents in some provinces, the rankings should be considered cautiously. Table 3 shows the results from regressions using these grouped variables. Background information about the respondent teachers were also collected and analyzed. The only data closely examined in this paper are those related to reactivity and incentives (the full survey is in Appendix 2).

The quantitative analyses are built around a conception of the dependent variable (Y) in this study (Diez, Barr & Cetinkaya-Rundel, 2016):

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

Y_i represents the dependent variable or reactivity level for 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 (all of the several incentives variables are named in Table 3); 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 ; and e_i is the residual of the regression.

The survey data were operationalized by assigning values to responses. Any practice or policy that seemed to promote or make easier teachers using LSA data was assigned a positive value. A practice or policy that was thought to make using these data more difficult was assigned a negative value (shown in Appendix 2). For reactivity variables, a similar process was used. Three choices were presented to respondents about the frequency of using specific strategies as a result of LSA data. Responses of 'always' were scored ±1, 'sometimes' was scored ±0.5, and 'never' was scored as 0 (TTT responses received negative values). These values are shown in alongside the survey in Appendix 2.

Each of the four lines of inquiry consisted of several survey questions which were aggregated for the preliminary analysis seen in Table 3. Each of the four aggregated scales were given equal weight in this regression in order to see which lines of inquiry had the most statistical impact on respondents' use of data.

Within each of the groupings, the questions asked were analyzed to see which of the independent variables had the greatest statistical impact in the use of LSA data to improve instructional practices (Table 4). Incentives variables are the main focus of this paper. Five questions were asked about incentives: i. Which jurisdiction (the school, school division, or provincial ministry, or none) had expectations that the data be used to improve instruction; 2. Which jurisdiction followed up on how LSA data were used in the classroom to improve instruction (the school, school division, or provincial ministry, or none); 3. Perceived pressure from LSAs was rated between the choices 'none', 'low' and 'high.' 4. Perceived stakes for the LSAs were rated on the same scale. 5. Respondents were rated on their awareness of results by answering whether their class, school or school division results were higher than average, lower than average, or average. The other options were 'I don't recall' or 'the results were not provided.'

Spearman's rank order correlation tests were done for these items (and items for all other lines of inquiry) to verify the validity of these numerical scales. Significant correlations between the independent variables indicated that the scales aligned, that many of the variables were relevant, and that values had been ascribed properly. The Spearman's correlations proved to be both significant and positive, but no distortion from multicollinearity was noted (Appendix, figure A1).

Cronbach's alpha was calculated to verify the internal consistency of scale variables in this table and in the analyses that follow. The alpha is often used as an index of reliability to check that items appear to measure the underlying construct (Jimerson, 2016; Santos, 1999). All reactivity scale items had alpha scores over 0.70 or more and the scale alpha was 0.742 indicating adequate

congruence of the items in the aggregated score. For the incentives scale, the alpha was 0.731 while the lowest item score was 0.649 (Appendix, figures A2 and A3). Figures for these and other statistical analyses are found in Appendix 1. Alpha values are subject to interpretation which is best left to the reader, noting that both scales have good consistency across items (Tavakol & Dennick, 2011).

Interviews

The interview data collection used a semi-structured format to follow up upon the themes from the survey instrument. This was in order to triangulate the qualitative and quantitative data (Hamilton et al., 2009; Jick, 1979). While interviews were done to explore the lines of inquiry in more depth, contradictory data were also unearthed. Knowing there is a wide range of opinion on the topic of LSA, the data sets are complementary in the sense that only through the exploration qualitative data can one hope to fully explain relationships found using quantitative methods (Onwuegbuzie & Leech, 2005).

Subjects for interviews were purposively selected to represent the instructional strategies self-reported by teachers (Flick, 2006). Selection was based on a stratified sampling at the high and low ends of the reactivity range (those teachers with high scores for TTC and also those with high scores for TTT were contacted). The sample was multilevel since it also included in-school and division-level administrators (McMillan & Schumacher, 2010). Only classroom teachers had completed the survey, but these in-school and division-level staff were interviewed in order to compare their responses to those of teachers.

The interview sample included teachers and administrators from across Canada and was made up of 13 classroom teachers, 10 in-school administrators, and four division-level staff. Respondents from such a small sample could not be expected to be representative of the survey population or to meet an external validity standard. They were chosen purposively to provide extra insight into the quantitative results (Flick, 2006). Interview data were not collected for generalization but rather to elaborate on specific topics.

Limitations

The data set from the survey was extensive but lacked some variability in locations where many schools and/or school divisions chose not to participate. The data were collected under strict confidentiality which meant the exclusion of school identifiers which might have made possible analysis using hierarchical linear modeling methods. HLM methods would be well suited to these data if the author had collected identifiers, and as a result these findings may be seen as less accurate. The response rate for the survey was reasonable but not high enough to dismiss the concern of nonresponse bias. While survey items were drawn from peer-reviewed and influential education studies, this is no guarantee of the construct validity. The interviews provided valuable insights into both policy and instructional choices, but were drawn from a small sample, and thus have their value limited to the insight they provide in conjunction with the survey data analysis.

Findings

Four Lines of Inquiry

The independent variables used in the survey were identified initially from the literature on large-scale assessment and included four main lines of inquiry operationalized for quantitative analysis purposes: (1) test data and design; (2) supports provided for teachers; (3) incentives for teachers to use these data; and (4) teachers' attitudes regarding the large-scale assessment. The first category asked about the types of items on the LSA, and how the data were returned to teachers. The

supports category asked what type and number of supports were made available to teachers. The third grouped variable consisted of questions about incentives which are seen below in Table 4. The attitudes variable asked teachers' opinions on how they thought the data might best be used. The complete survey is found in Appendix 2 and makes clear the categorized variable groupings.

Table 3 presents the lines of inquiry with beta (standardized) coefficients in order to show the relative strength of scale variables in an unbiased comparison. Coefficients are from multivariate regressions for the nationwide dataset, which allows one to see which scale variables had the most significant effects on different types of reactivity. For detailed analysis of lines of inquiry other than incentives, see Copp (2016a, 2016b, 2017).

Table 2
Reactivity scores categorized into TTC, TTT and net effects[6]

	Prov.	Avg. score	SD	Prov.	Avg. score	SD	Prov.	Avg. score	SD		
Rank		TTC			TTT			Net Reactivity		Prov.	*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		-3.09	1.24		-0.56	1.37		418

Note: TTT scores are negatively valued. Short forms are used for the provinces: Alberta (AB); British Colombia (BC); Manitoba (MB); New Brunswick (NB); Newfoundland and Labrador (NL); Nova Scotia (NS); Ontario (ON); Prince Edward Island (PEI; Quebec (QC) and Saskatchewan (SK).

The variables with the greatest correlation with TTC, based on beta scores, are the attitudes variables (Copp, 2016a). They have no significant impact on TTT effects. Similarly, supports variables have a significant correlation with TTC but none on TTT (Copp, 2016c). Tests and data variables have no significant impacts on either TTC or TTT. Most relevant to this paper, though, are incentives variables which have a uniquely significant impact on both TTC and TTT. With three significant scale variables in play, the amount of variance in TTC explained by these lines of inquiry is a relatively high 20%. With only one significant scale variable working on TTT effects, the adjusted R^2 value falls to only 3%. It can be seen that incentives variables are significant in their correlation to LSA-based instructional change.

Incentives Variables

[6] Adapted from Copp, 2016b, p. 11.

There were five independent variables under the grouping 'incentives' examined in this study. The most significant of these (as seen in Table 4) was 'perceived pressure.' It shows highly significant correlations with both TTC and TTT effects before and after the addition of provincial dummy variables (these being included to account somewhat for variations between testing programs in different jurisdictions). Perceived pressure was also the only variable that proved significant at all on the TTT side of the table after provincial dummies were included in the regression model.

Table 3
Four lines of inquiry from the survey correlated with instructional changes (TTC and TTT).[7]

		TTC		TTT	
		Coefficient	SE	Coefficient	SE
Tests and data		0.00	0.002	0.002	0.004
		(0.004)	(0.004)	(0.005)	(0.005)
Supports		0.014**	0.0124*	0.009	0.007
		(0.005)	(0.005)	(0.006)	(0.006)
Incentives		0.083***	0.083***	-0.073**	-0.063**
		(0.019)	(0.021)	(0.023)	(0.023)
Attitudes		0.015**	0.013**	-0.002	-0.004
		(0.005)	(0.005)	(0.005)	(0.005)
Provincial dummies	AB	0.109		AB	0.283
		(0.299)			(0.397)
	BC	-0.074		MB	-0.455
		(0.381)			(0.421)
	NB	0.048		NB	0.397
		(0.295)			(0.402)
	NL	0.088		NL	-0.128
		(0.341)			(0.440)
	NS	-0.298		NS	1.460**
		(0.310)			(0.402)
	ON	-0.004		ON	0.718
		(0.307)			(0.413)
	PEI	-0.702		PEI	0.069
		(0.301)			(0.395)
	QC	0.018		QC	-0.052
		(0.322)			(0.430)
	SK	-0.629		SK	1.099*
		(0.355)			(0.456)
Constant		0.974**	1.147**	-3.232***	-3.670***
		(0.345)	(0.425)	(0.395)	(0.454)
N		228	228	230	230
Adj. R^2		0.199	0.231	0.033	0.148

[7] Adapted from Copp, 2016a, p. 12.

Note: SE in parentheses; * $p<0.05$, ** $p<0.01$, *** $p<0.001$; TTT effects are negatively valued. Short forms are used for the provinces: Alberta (AB); British Colombia (BC); Manitoba (MB); New Brunswick (NB); Newfoundland and Labrador (NL); Nova Scotia (NS); Ontario (ON); Prince Edward Island (PEI; Quebec (QC) and Saskatchewan (SK).

Table 4
Incentives variables correlated with the use of data to inform instructional changes (TTC and TTT)

	TTC			**TTT**		
	Coefficient	SE	*β*	Coefficient	SE	*β*
Expectation to use data	0.062	0.032	0.043	0.056	0.012	0.015
	(0.044)	(0.047)		(0.050)	(0.052)	
Follow up on data use	0.120*	0.127*	0.137	-0.002	0.019	0.019
	(0.053)	(0.053)		(0.061)	(0.059)	
Results-awareness	0.039	0.044	0.081	-0.070*	-0.044	-0.076
	(0.028)	(0.028)		(0.032)	(0.031)	
Perceived pressure	0.778***	0.764***	0.22	-0.787***	-0.652**	-0.177
	(0.195)	(0.194)		(0.220)	(0.214)	
Perceived stakes	0.356*	0.374*	0.118	-0.207	-0.019	-0.006
	(0.173)	(0.173)		(0.196)	(0.191)	
	AB	0.290		AB	-0.271	
		(0.255)			(0.278)	
	BC	-0.456		MB	0.304	
		(0.258)			(0.281)	
	NB	0.094		NB	0.058	
		(0.254)			(0.278)	
	NL	-0.024		NL	-0.225	
		(0.289)			(0.314)	
	NS	0.061		NS	0.978***	
		(0.263)			(0.285)	
	ON	-0.010		ON	0.457	
		(0.275)			(0.299)	
	PEI	-0.592*		PEI	-0.310	
		(0.256)			(0.281)	
	QC	0.093		QC	-0.235	
		(0.279)			(0.303)	
	SK	-0.490		SK	0.676*	
		(0.304)			(.0334)	
Constant	1.755***	1.870***		-2.511***	-2.795***	
	(0.131)	(0.209)		(0.146)	(0.232)	
N	343	343		344	344	
Adjusted R^2	0.163	0.205		0.061	0.147	

*Note: * p<0.05, ** p<0.01, *** p<0.001; β (beta) scores show standardized values; TTT effects are negatively valued. Short forms for provinces: Alberta (AB); British Colombia (BC); Manitoba (MB); New Brunswick (NB); Newfoundland and Labrador (NL); Nova Scotia (NS); Ontario (ON); Prince Edward Island (PEI; Quebec (QC) and Saskatchewan (SK).*

The other variables that indicated significant results were 'follow up data use' and 'perceived stakes' these having significant correlations with only TTC effects. Each of these variables will be considered in turn alongside qualitative (interview) data used to either support or challenge the statistical correlations shown here.

Perceived pressure. The perceived pressure variable was highly significant in terms of TTC and TTT effects on teachers. There is evidence here to support the idea that incentives used to increase the pressure on teachers to use LSA data are quite effective at this task. It is less encouraging that perceived pressure does not significantly increase the chances that teachers are using the data in the manner expected by provincial education ministries or the STF Code of Professional Conduct (2015). Incentives that increase pressure on teachers increase both TTT and TTC effects, and make less certain the educational effectiveness of the instructional changes to which their application and implementation appears to contribute (Appendix, figure A8).

Interview respondents frequently commented on the high levels of pressure teachers felt when teaching classes that administered LSAs.

> As an administrator I try to assure them [teachers] that this is part of your teaching, this is all doing the things you do normally. I still think there is a stress there, that their kids, if their kids haven't met the standards there is obviously a stress there that they are not doing their job. (MB, Elementary school principal, female)

> We talk about the journey being a 10 year journey and so if you're missing some essential foundational learnings, and umm, then you have a Math assessment at the end of grade 9 and you haven't really achieved, I think our grade 9 teacher feels like, you know, a lot of pressure for her… We keep trying to remind her that they didn't 'not learn it' over one year, they 'not learned it' over a number of years. (PEI, K-9 school principal, female)

> And that's the unfortunate side-effect of a test culture, right? Like if you put all of the worth on the test, then of course your energies are going to be focused on the test which is not where there should be focused. (AB, High school Math teacher, female)

It was true that the source of the pressure was sometimes attributed to the public pressure for accountability made explicit through the publication of assessment results and implicit expectations from school or senior administration to perform well. More commonly, though, respondents spoke about professional responsibility being a source of pressure.

> A teacher is a teacher because of their intrinsic drive. I don't think they need any external forces pushing them to be better. (NB, Middle years homeroom teacher, female)

> You know I feel under a lot of pressure, but you know, it's self-induced, really, to make sure they do well, and the students are under that same pressure… You see across the hallway there are students over there that do course that don't have provincial exams and they seem to really be enjoying their courses… But in my room, you know, I don't want anyone knocking on the door from 9:00 until 10:00 - we're doing problems and nobody moves. (NL, High school Science teacher, male)

It was difficult in most cases to fault the teachers for their conduct knowing that the accountability systems in which they acted were not of their creation and were not designed according to their wishes. The understanding of this reality was especially clear in the minds of division-level administrators.

> It makes up such big part of kids' marks that in lots of ways it determines the kids' marks and those marks are, for most kids, only relevant when it comes to applying for post-secondary and applying for scholarships. And then you are looking at this kid gets in because they have and 86 and this kid doesn't because they have an 85. (AB, High school Math teacher, female)

> I don't think it is appropriate for a teachers to get old FSA exams 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)

> And I think [LSAs] do provide unnecessary pressure on teachers and I think they are misrepresented definitely in the media. And so, of course, people who don't understand and they just hear a test score or they just see, you know, Nova Scotia's results in literacy and math are low, and therefore our school system is poor and our teachers are not teaching effectively. Well, that is just not true. (NS, Division staff, female)

> The OSSLT is distinguished from the other 3 tests, namely grade 3, grade 6, and grade 9, because it is a graduation requirement. So, you know, it is not ethical as an educator, no matter where you stood on the fence with this, to not do your absolute best to allow students to be successful. (ON, Division staff, male)

It was clear from interviews (which also proved supportive of the quantitative findings) that pressure was the single most important driver of teachers' response to LSA testing and the release of results data. Whether internally or externally imposed, pressure caused significant stress to teachers and in some instances affected their choices about which grades or subjects they wanted to teach.

Follow up on data use. Following up on data use is an important aspect of policies being faithfully implemented (i.e. as policy makers had intended). The results from Table 2 appear to bear out this conclusion, yet the correlation is only weakly significant. Teachers had vastly differing experiences with this kind of oversight, some saying it was not evident, while others indicated that it was both prominent and important. The follow up that was noted by respondents came from different levels of administration, but school-level oversight was most frequently cited (Appendix, figures A5 and A6).

> It would vary widely depending on the administrative team at a given school. I have worked with principals for whom diploma results [LSA scores] would be 'be-all and end-all' and for whom every kid would be exempted from diploma exams if that were an option… it is highly inconsistent. (AB, High school Math teacher, female)

> The current administration is content with the fact that we are talking with each other and we talk to them and we discuss the concerns that we have. Everything tends to be

> a proactive approach. I think that's one of the values I share with my current administration. (AB, High school Science teacher, male)

> No [division-level follow up], none. None whatsoever. We choose to do [common assessments] because that's the only way we can grow as a school, to monitor the progress of our students. And it is important for us to have that data… to try to become a more data-driven school and show how, ahh, how we can improve as a result of using the data from previous years. (NB, High school principal, male)

For other schools, it was noted that the leadership to follow up on data use was not modeled from the division level or from the school administration. Some respondents painted the picture of schools and teachers sailing in a ship without a rudder in terms of data use.

> The division seems to be scattered and all over the place. . . We don't see anyone from the board. . . And, you can see it kind of falling away now - they don't ask us for reports anymore, 'Yeah, go ahead and do that…" (NL, High school Science teacher, male)

> Well, I know as an administrator you're certainly in and checking and following along with what [teachers] should be doing, but I don't know there would be necessarily a prescription for a specific teacher to follow this type of direction. And I would not be made aware of it from a board or department level that, 'Mr. Smith needs to improve on this portion of his lesson and it is important for you… to go in and watch and see that this is happening.' (PEI, High school Math teacher, male

> I don't think that expectation was ever given to us on what to do with them. It was sort of like, I felt like, oh, you have to give it to see where students are, and then there was no follow up from anyone. It was just, this is what it is… and that was it. (SK, Middle years homeroom teacher, female)

Considering the wide range of responses regarding following up on data use it is no surprise that the correlation is only weakly significant. A stronger correlation to reactivity effects would be dependent upon both more teachers indicating that some kind of follow up on their data use practices was done and also upon policy documents making it very clear to school and senior administration that they have this particular role to play to ensure common, faithful implementation.

Perceived stakes. High stakes testing is quite different in the United States as compared to Canadian LSAs. The stakes in Canada are directly applied only to students, and indirectly to teachers and schools. The evidence on the efficacy of stakes as an incentive is inconclusive, but has been more closely tied to TTT effects than the TTC variety.

Many teachers made note of the lack of professional stakes applied to teachers based on their students' LSA scores (Appendix, figure A9).

> I can't recall anyone having paid a professional price for bad results. So, yes, it is high stakes, teachers are apprehensive, but I think that it has been enough of a routine that teachers manage that. (AB, High school Science teacher, male)

> As an administrator the conversation about, 'you need to improve those scores or else', kind of thing, has never happened. Yeah, we don't go that road. (MB, Elementary school principal, female)

> I see the newer teachers worried like, 'Am I going to get to teach this class again?' Because there is still a perceived hierarchy in the courses you get to teach. (AB, High school Math teacher, female)

It is also the case that many respondents saw the stakes they perceive which are applied from higher jurisdictional levels and can, for good or for ill, have an effect on their professional development and/or their teaching practices.

> Yeah, just from knowing [the provincial exam] is coming up, you know, even weighing my decision about whether or not I wanted to loop with my kids or not. I wanted to loop but I knew that [the exam] was at the end of the year, I was going to have to face that exam, those exams. (QC, Middle years homeroom teacher, female)

> What happens in literacy is entirely, well, not entirely, but the teacher has the plasticity around what happens in that time frame. The same way in their math block… It is less about robbing Peter to pay Paul and more about looking at much broader strokes of how we approach literacy and numeracy. (ON, Division staff, male)

> Whereas the math [assessment], yes, absolutely. I know 100% that it is used in our board to compare teachers to teachers. (ON, High school English consultant, female)

Some concern was raised about how the data can help create perceptions that are not entirely accurate. A snapshot was not seen as a very good metric for measuring teacher effectiveness.

> I have to be honest, I only look at our own school, you know. And generally I think we are getting better, or the marks are getting better. Whether that means [teaching is] getting better, I don't know. . . Sometimes I think it is just the group of kids you have. (PEI, Elementary homeroom teacher, female)

> When you are comparing from year to year and using that as data for improvement it is kind of like trying to judge the population trends of height, for example, by watching people pass by a window. A tall person passes, then a medium person passes, then a short person passes. If you use that process to do your data gathering, then you are going to conclude that people are getting shorter. (QC, High school English teacher, male)

Since stakes for teachers are applied at much lower levels than it is in some other high stakes environments (many interview respondents referenced the US education system), this proved to be a less significant variable than the pressure which is more clearly apparent from students, parents and the public at large.

Other variables. Some of the variables examined did not prove significant despite their prominence in the LSA literature. Being aware of the results of LSAs is clearly the entry point to using the data (Datnow, Park & Kennedy-Lewis, 2012; Wideman, 2002), but this variable did not prove significant for either instructional change strategy barring the weak correlation to TTT before the addition of provincial dummies. Teachers reported their levels of awareness were quite dependent on the attitudes of school and divisional leadership (Appendix, figure A7).

> Our administration is very supportive and do not use the data for anything within our school but provincially the data is used to rank schools in the whole province. (BC, Elementary homeroom teacher, female)

> Now... all of our resources are supposed to be allocated and all of our teaching is supposed to be data-driven. Right? So we are all about collecting the data and we've got testing up the wazoo and stuff. So there is on some level, there is an expectation that we are trying to improve our results every year. (PEI, Elementary homeroom teacher, female)

Another aspect of implementation literature examined in this study was the effectiveness of making expectations explicit and clear. Even when expectations were relayed to teachers, no correlation with reactivity effects appears in the quantitative data. The qualitative data indicate that teachers have very different experiences regarding this variable (Appendix, figure A4).

> It is district-driven. We have a school improvement plan that we have to do every year and basically it is instructional-based… So we do that at the beginning of the year. When we go in in August we're gonna talk about what are we looking at that we really want to focus our instruction on. So using the data, the district requires that we fill out an SIP. (NB, Middle years homeroom teacher, female)

> I think unless there is involvement of the administration in the, you know, in a literacy culture, in a standardized test culture in the school, all classroom teachers can close their doors and do whatever they want. (ON, High school English consultant, female)

It appears that expectations, much like all the other variables in this section, are relayed in very different ways to different teachers and in some cases, they are not transmitted at all. Since policy expectations across Canada differ, it may be considered unfair to judge the practical effects of the implementation for such different LSA programs. Yet we should remember the fact that the policy goals of LSA are strikingly similar across all 10 provinces, and that the improvement of instruction (referred to in this paper as teaching [to] the curriculum) is the ultimate goal of all provincial education ministries. In sum, these responses show that the uneven and unclear implementation practices reported by interview subjects align well (triangulate) with the survey results to illuminate practices that: a) do not always align to stated policy goals; b) that have unintended consequences on instructional practices; and c) which divert the attention, financial resources and the time of educational professionals and leaders in this pursuit.

Discussion

In an effort to both improve the academic results of students in large-scale assessments and to improve the quality of teaching in Canadian public schools, LSAs are used in all 10 Canadian provinces as a means of reaching these goals. Unfortunately LSAs have not proven to be a uniformly effective means of promoting positive instructional changes. This study examined the use of LSAs and the incentives built into assessment policies in the context of both the policy goals stated by provincial ministries and the instructional methods demonstrated as effective in the STF Code of Professional Conduct (2015). It has been seen that policy-based incentives do correlate with more reactivity, but it is predominantly in the form of teaching to the test, which is in line with neither ministry goals nor the principles of the STF Code. This being established, there are some lessons for policy makers to draw.

'Incentives Work'

Incentives variables do prove to be quite effective at inspiring instructional change. Of the scale variables examined in the larger study upon which this paper is based, the incentives scale alone showed highly significant results in terms of teaching (to) the curriculum effects as well as teaching to the test effects. Incentives are clearly very effective in terms of increasing the level of reactivity in teachers. Yet while TTC is the stated goal of both the provincial ministries and the STF Code, TTT is not, and yet it has an equally significant relationship with policy incentives. Recent research bears out the fact that TTT is a common and less desirable consequence of the single-minded pursuit of higher test scores (Hill, Mellon, Laker & Goddard, 2016).

So while it is seen that incentives are effective at promoting instructional change, it should be added that the changes which result from policy incentives do not necessarily help ministries achieve instructional improvement goals. Thus the use of policy-based incentives is at best a questionable method of trying to achieve either better teaching or better learning.

High Stakes

The correlation found in the literature between high stakes testing and high pressure testing environments is an important one to consider in this study (Finnigan & Gross, 2007; Luna & Turner, 2001; Madaus, 1988). Of the 10 provinces in Canada, seven currently employ some form of high stakes exam whether it be a minimum competency exam or one that is factored into grades for summative purposes. Since graduation, university admissions, and scholarships all depend on these results, teachers feel great pressure from students, parents and administrators to produce solid results. There is an apparent correlation between giving these high stakes, high-pressure exams and greater reported reactivity effects. Table 3 lists the provinces in terms of reactivity effects.

The three provinces which do not give high stakes high school exams are Nova Scotia, Prince Edward Island, and Saskatchewan. These provinces appear near the bottom of the TTT scale and the middle of the TTC scale. The same group also appears near the top of the net reactivity scale which (by adding the positive values of TTC scores to the negative values applied to TTT) gauges the balance between the self-reported uses of these strategies. If LSA policies (including incentives) were acting as proposed, the net reactivity scale should be heavily tilted into the positive range. We can see that there is less reactivity generally in these three provinces, but that they have the highest degree of balance between TTC and TTT effects (a perfect balance would produce a net reactivity score of 0).

In terms of policy, the use of high stakes or minimum competency exams is evident in the majority of Canadian educational jurisdictions. Seeing that all the provinces that employ such assessment tools also score higher in net reactivity than the three provinces that do not use them should give ministry officials reason to reconsider their use. If the use of high stakes exams is correlated to TTT effects, it is of no help in reaching stated goals for improved teaching.

Pressure

The most significant of the variables within the incentives scale is perceived pressure. Teachers reported feeling high levels of pressure despite the Canadian context of LSA showing notably less professional consequence for educators as compared to US policy models (Finnigan & Gross, 2007; Koretz, 2009). It seems obvious that the policy-based administration of pressure needs to be considered in terms of the effects it has on instructional practices. Where reactivity effects contradict fundamental policy goals both expectations and incentives should be re-evaluated.

It is difficult for policy makers to control a fuzzy variable dealing with 'perceptions', but it should be emphasized that pressure is applied inadvertently, explicitly, or from educational

stakeholders (such as the community, students or parents). This pressure is strongly correlated to the use of TTT strategies. This 'test anxiety for teachers' is not conducive to meeting ministry policy targets.

It may not seem particularly constructive to point out only things that policy makers and education ministry officials should *not* do. In fields of endeavour where there is the potential of harmful effects resulting from our actions, we could keep in mind that medicine's ancient Hippocratic Oath begins with an admonition to 'do no harm.' This is what we would now call the precautionary principle – the practice of exercising caution when understanding is incomplete. When incentives are examined as a policy tool, the limiting of harm must be considered a primary goal since there is so very little to say about the positive influence that policy incentives have on pedagogical practice.

Conclusion

The inclusion of incentives in provincial assessment policies across Canada has shown to be an effective means to promote the use of data in classrooms: teachers are significantly reactive to incentives. If this study did not differentiate between TTT and TTC reactivity effects, this would be seen as further proof that incentives are a suitable and appropriate lever. The main driver of this reactivity appears to be perceived pressure coming from test preparation, the high stress and sometimes secretive methods of test administration, and finally the public release of results. Explicit policy goals and guidelines for teacher conduct do not advocate teaching to the test as a preferred or effective instructional strategy, yet the LSA policies are equally likely to have this effect as they are to lead to teaching (to) the curriculum. An educational experiment with such unpredictable outcomes can hardly be called a success. Only when it can said with some certainty what effect the pressure from incentives will have on teachers can policy makers claim to have achieved some measure of successful implementation.

Based on these results the obvious conclusion is that for too long the model based on incentives, themselves inspired by capitalist economic models, have been used to guide policy in the public sphere. This is despite growing evidence that even in business incentives have unintended consequences (see, for example, Ariely, Gneezy, Loewenstein & Mazar, 2009). It seems equally apparent that policies built on the presumption of 'improvement' should be tested against objective criteria that accurately gauge their practical effects. This program evaluation study has asked that basic question: whether assessment policy is actually meeting the outcomes originally put forward in ministry literature.

The results of this research go some way to showing that incentive-based policies, especially those rooted on high-pressure testing, do not result in better instructional practices since TTC improvements are eclipsed by changes to TTT methods (as in Table 2). Whether or not increasing LSA scores indicate improved teaching (and this is an important question not explored here), high quality instruction is the stated and laudable goal of all education ministries in Canada. In order to achieve it, more objective examinations of the practical results of LSA must be conducted and examined with a critical eye. Being the basis of a large research study, the author has several papers built on this very premise which all examine the practical effects of LSAs on teaching in Canadian public schools using the reactivity model. It is hoped that these papers and the ongoing research on assessment will help policy makers avoid the possible unintended consequences of incentives-based programs and both devise and implement LSAs that support and promote the more effective instructional strategies which help teachers teach (to) the curriculum.

References

Abrams, L. M. (2004). *Teachers' views on high-stakes testing: Implications for the classroom*. Education Policy Studies Laboratory. EPSL Working Paper EPSL-0401-104-EPRU. Retrieved Mar. 30, 2013, from http://epsl.asu.edu/epru/documents/EPSL-0401-104-EPRU.pdf

Allen, J. R. (2002). *Value-for-money in Saskatchewan K – 12 educational expenditures*. Saskatchewan Institute of Public Policy. SIPP Public Policy Paper No. 10. Retrieved Sep. 22, 2017, from https://www.researchgate.net/profile/John_Allan3/publication/251251045_Value-for-Money_in_Saskatchewan_K-12_Educational_Expenditures/links/54dcfed80cf28a3d93f88907/Value-for-Money-in-Saskatchewan-K-12-Educational-Expenditures.pdf

Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement, 26*(1), 32-56.

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Economic Policy Research Unit. EPRU Working paper. Retrieved Nov. 26, 2013, from http://nepc.colorado.edu/files/EPSL-0211-125-EPRU.pdf

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. The *Review of Economic Studies, 76*(2), 451-469.

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

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77-86.

Bauer, S. C. (2000). Should achievement tests be used to judge school quality? *Education Policy Analysis Archives, 8*(46), 1–18.

Ben Jaafar, S., & Earl, L. (2008). Comparing performance-based accountability models: A Canadian example. *Canadian Journal of Education, 31*(3), 697–725.

Bishop, J. H., & Wößmann, L. (2004). Institutional effects in a simple model of educational production. *Education Economics, 12*(1), 17–38.

Blaikie, N. (2000). *Designing social research*. Cambridge: Polity Press.

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), 45-68.

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, No. 71. OECD Publishing. http://dx.doi.org/10.1787/5k9fdfqffr28-en

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. http://dx.doi.org/10.1080/0969594042000304609

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297-312.

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability, 21*(1), 47-65.

Chambers, S., & Tate, K. (2015). Value-added measures in restructuring urban schools. *Education and Urban Society, 47*(6), 723–738. doi:10.1177/0013124513508582

Cizek, G. J. (2005). More unintended consequences of high-stakes testing. *Educational Measurement:*

Issues and Practice, 20(4), 19-27.
Copp, D. T. (2015). *Teacher-based reactivity to provincial large-scale assessment in Canada*. Maastricht, the Netherlands: Boekenplan.
Copp, D. T. (2016a). 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). http://dx.doi.org/10.14507/epaa.24.2522
Copp, D. T. (2016b). Teaching to the test: a mixed methods study of instructional change from large-scale testing in Canadian schools. *Assessment in Education: Principles, Policy & Practice.* http://dx.doi.org/10.1080/0969594X.2016.1244042
Copp, D. T. (2017). Accountability testing in Canada: Teacher perspectives on assessment programs. Manuscript submitted for peer review.
Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be?: The use of value-added measures of teacher effectiveness in policy and practice.* New York, NY: Annenberg Institute for School Reform.
Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*(4), 464–94.
Cullen, J. B., & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system.* National Bureau of Economic Research. NBER Working Paper Series 12286. Retrieved Apr. 1, 2013 from http://www.nber.org/papers/w12286
Datnow, A., Park, V. & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *Journal of Education for Students Placed at Risk (JESPAR), 17*(4), 247-265. http://dx.doi.org/10.1080/10824669.2012.718944
Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297.
Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2016). *OpenIntro statistics* (3rd ed.). Lexington, KY: CreateSpace.
Ehren, M. C., & Shackleton, N. (2016). Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools. *Educational Assessment, Evaluation and Accountability, 28*(4), 299-321.
Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology, 113*, 1-40.
Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal, 44*(3), 594–630. http://dx.doi.org/10.3102/0002831207306767
Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*(2), 95–113.
Flick, U. (2006). *An introduction to qualitative research* (3rd ed.). London: Sage Publications.
Fullan, M. (2009). Large-scale reform comes of age. *Journal of Educational Change, 10*(2-3), 101–113. http://dx.doi.org/10.1007/s10833-009-9108-z
Goertz, M. E., Oláh, L. N., & Riggan, M. (2009*). From testing to teaching: The use of interim assessments in classroom instruction*. CPRE Research Reports. Retrieved from http://repository.upenn.edu/cpre_researchreports/58
Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments*. RAND Education. RAND Working Paper WR-374-EDU. Retrieved Sep. 22, 2017 ,from https://www.researchgate.net/profile/Laura_Hamilton6/publication/265100351_Instructio

nal_Practices_Related_to_Standards_and_Assessments/links/544931650cf2f638808109f6.pdf

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., . . . Steele, J. L. (2009). *Using student achievement data to support instructional decision making.* United States Department of Education. Retrieved from http://repository.upenn.edu/gse_pubs/279

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*(1), 131–157.

Hargreaves, A., & Shirley, D. (2011). *The far side of educational reform.* Canadian Teacher's Federation Brief. Retrieved Sep. 22, 2017, from https://www.ctf-fce.ca/Research-Library/Report_EducationReform2012_EN_web.pdf

Hill, A., Mellon, L., Laker, B., & Goddard, J. (2016). The one type of leader who can turn around a failing school. *Harvard Business Review, 20*.

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation* (pp. 163-189). Greenwich, CT: Information Age Publishing.

Hood, C. (2006). Gaming in targetworld: The targets approach to managing British public services. *Public Administration Review, 66*(4), 515-521. http://dx.doi.org/10.1111/j.15406210.2006.00612.x

Huber, S. G., & Skedsmo, G. (2016). Data use—a key to improve teaching and learning? *Educational Assessment, Evaluation and Accountability, 28*(1), 1-3.

Jacob, B. A. (2002, Jun.). Test-based accountability and student achievement gains: Theory and evidence. Taking account of accountability: Assessing politics and policy. Lecture conducted from John F. Kennedy School of Government, Harvard University, Cambridge MA.

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

Jimerson, J. B. (2016). How are we approaching data-informed practice? Development of the survey of data use and professional learning. *Educational Assessment, Evaluation and Accountability, 28*(1), 61-87.

Klinger, D. A., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy, 76*, 1-34.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review, 28*(6), 682–692.

Koretz, D. (2009, Dec.). Implications of current policy for educational measurement. Lecture conducted from the Center for K-12 Assessment & Performance Management, Princeton, NJ.

Levin, B., Glaze, A., & Fullan, M. (2008). Results without rancor or ranking: Ontario"s success story. *Phi Delta Kappan, 90*(4), 273-280.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3–13. http://dx.doi.org/10.3102/0013189X029002004

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal, 91*(1), 79-87.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46. http://dx.doi.org/10.1080/01619568809538611

Maharaj, S. (2017, Sep. 20). Toronto District School Board reacts to declining EQAO math scores. *Global News*. Retrieved Sep. 22, 2017, from https://globalnews.ca/news/3759761/tdsb-eqao-math-scores/

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate:

Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education, 60*, 366-376. http://dx.doi.org/10.1016/j.tate.2016.07.011

Marsh, J. A., Farrell, C. C., & Bertrand, M. (2014). Trickle-down accountability: How middle school teachers engage students in data use. *Educational Policy, 30*(2), 243-280.

McMillan, J. H., & Schumacher, S. (2010). *Research in education: Evidence-based inquiry* (7th ed.). Boston: Pearson.

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). http://dx.doi.org/10.14507/epaa.v11n3.2003

Møller, J. (2008). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational Change, 10*(1), 37–46. http://dx.doi.org/10.1007/s10833-008-9078-6

Morgan, C. (2009, May 27). Transnational governance: The case of the OECD PISA. Lecture conducted from the Canadian Political Science Association (CPSA), Ottawa, Canada.

Morris, A. (2011). *Student standardised testing: Current practices in OECD countries and a literature review*. OECD Education Working Papers, No. 65. OECD Publishing. http://dx.doi.org/10.1787/5kg3rp9qbnr6-en

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

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In Brown, G. T., & Harris, L. R. (Eds.), (pp. 40-56). *Handbook of Human and Social Conditions in Assessment*. New York, NY: Routledge.

Oláh, L. N., Lawrence, N., & Riggan, M. (2008, Mar. 27). Learning to learn from benchmark assessment data: How teachers analyze results. Lecture conducted from the American Educational Research Association, New York, USA.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly, 70*(5), 737–758. http://dx.doi.org/10.1093/poq/nfl038

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. http://dx.doi.org/10.1080/13645570500402447

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

Rockoff, J. E. (2003). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247–252. http://dx.doi.org/10.1257/0002828041302244

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher, 45*(4), 252-257. http://dx.doi.org/10.3102/0013189X16649961.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension, 37*(2), 1-5.

Saskatchewan Teachers' Federation. (2015). STF governance handbook: Code of professional competence. Retrieved July 12, 2016, from http://www.stf.sk.ca/sites/default/files/governance_handbook_bylaw_7_2.pdf

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education, 26*(3), 482–496. http://dx.doi.org/10.1016/j.tate.2009.06.007

Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2016). Data teams for school improvement.

School Effectiveness and School Improvement, 27(2), 228-254. http://dx.doi.org/10.1080/09243453.2015.1056192

Scott, S., Webber, C. F., Aitkin, N., & Lupart, J. (2011). Developing teachers' knowledge, beliefs, and expertise: Findings from the Alberta student assessment study. *The Educational Forum, 75*(2). http://dx.doi.org/10.1080/00131725.2011.552594

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14. http://dx.doi.org/10.3102/0013189X029007004

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

Spillane, J. P., Diamond, J. B., Burch, P., Hallett, T., Jita, L., & Zoltners, J. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy, 16*(5), 731–762. http://dx.doi.org/10.1177/089590402237311

Statistics Canada. (2007). Education indicators in Canada: Report of the Pan-Canadian Education Indicators program 2007. Retrieved Apr. 22, 2013, from http://publications.gc.ca/collections/collection_2007/statcan/81-582-X/81-582-XIE2007001.pdf

Stoilescu, D., McDougall, D., & Egodawatte, G. (2016). Teachers' views of the challenges of teaching grade 9 applied mathematics in Toronto schools. *Educational Research for Policy and Practice, 15*(2), 83-97. http://dx.doi.org/10.1007/s10671-015-9178-z

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2,* 53-55. http://dx.doi.org/10.5116/ijme.4dfb.8dfd

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. http://dx.doi.org/10.1080/15309576.2002.11643661

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35*.

Volante, L., & Ben Jaafar, S. (2008). Profiles of education assessment systems worldwide. *Assessment in Education: Principles, Policy & Practice, 15*(2), 201-210. http://dx.doi.org/10.1080/09695940802164226

Volante, L., & Cherubini, L. (2010). Understanding the connections between large-scale assessment and school improvement planning. *Canadian Journal of Educational Administration and Policy*, 115.

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). http://dx.doi.org/10.14507/epaa.v20n25.2012

Webb, P. T. (2006). The choreography of accountability. *Journal of Education Policy, 21*(2), 201–214. http://dx.doi.org/10.1080/02680930500500450

Weinbaum, E. H. (2009). *Learning about assessment: An evaluation of a ten-state effort to build assessment capacity in high schools*. Consortium for Policy Research in Education (CPRE). CPRE Research Report #RR-61. Retrieved February 16, 2014, from http://cpre.org/sites/default/files/researchreport/827_cpreten-stateassessmentweb-copy.pdf

Wideman, R. (2002). Using action research and provincial test results to improve student learning. *IEJLL: International Electronic Journal for Leadership in Learning, 6*(20).

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist, 45*(2), 107–122.

Woessmann, L. (2001). Why students in some countries do better: International evidence on the importance of education policy. *Education Matters, 1*(2), 67–74.

Wößmann, L. (2003). *Central exams as the "currency" of school systems: International evidence on the complementarity of school autonomy and central exams*. CESifo DICE Report, *1*(4), 46-56.

Young, V. M. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education, 112*(4), 521-548.

Appendix 1

The charts found in this appendix are intended to provide more information regarding the study parameters and results. Where the preceding paper does not provide a sufficient amount of detail, often in the interest of space, the appendices will hopefully make these points clear. What follows is: a) Spearman's rank order correlation tests for incentives scale items; b) Cronbach's alpha scores for reactivity scale and incentives scale items; c) descriptive statistical data for incentives variables; and in Appendix 2, d) the survey instrument questions with the ascribed numerical values in parenthesis.

	(1)	(2)	(3)	(4)	(5)
(1) Expectation to use data	1.000				
(2) Follow up on data use	0.472*	1.000			
(3) Perceived pressure	0.201*	0.199*	1.000		
(4) Results awareness	0.311*	0.286*	0.138*	1.000	
(5) Perceived stakes	0.261*	0.212*	0.158*	0.401*	1.000

* p<0.05, ** p<0.01

Figure A1: Spearman's rank order correlation test for incentive scale items.

Item	*n*	Item-test correlation	Item-rest correlation	Average inter-item covariance	Alpha
Look for PD	413	0.530	0.389	0.029	0.722
Request resources	408	0.574	0.429	0.028	0.717
Work with other teachers	410	0.496	0.335	0.029	0.730
Cover a wider range	409	0.473	0.287	0.029	0.741
Offer study groups	408	0.477	0.310	0.029	0.735
Cover tested content well	414	0.649	0.543	0.027	0.704
Teach test-taking strategies	410	0.585	0.455	0.028	0.714
Use similar classroom tests	412	0.635	0.521	0.027	0.705
Focus on tested subjects	409	0.523	0.352	0.028	0.730
Review old exams	414	0.619	0.474	0.027	0.709
Test scale				0.028	0.742

Figure A2: Cronbach's alpha scores for reactivity scale items.
All scores are within reasonable bounds, whereas alphas for TTT and TTC items separately are lower as a result of the small number of included items.

Item	*n*	Item-test correlation	Item-rest correlation	Average inter-item covariance	Alpha
Expectation to use data	1071	0.813	0.597	0.141	0.697
Follow up on data use	1071	0.612	0.397	0.166	0.643
Perceived pressure	411	0.441	0.36	0.231	0.732
Aware of class results	369	0.582	0.415	0.187	0.695
Aware of school results	363	0.666	0.526	0.173	0.672
Aware of division result	359	0.585	0.381	0.182	0.700
Perceived stakes	410	0.387	0.291	0.233	0.735
Test scale				0.189	0.731

Figure A3: Cronbach's alpha for incentive scale items.
The higher n for expectation and follow up data is a result of the inclusion of all surveyed teachers in these item responses, whether or not they gave LSAs in their classrooms. Comparisons of the two groups are found below in figures A9 and A11.

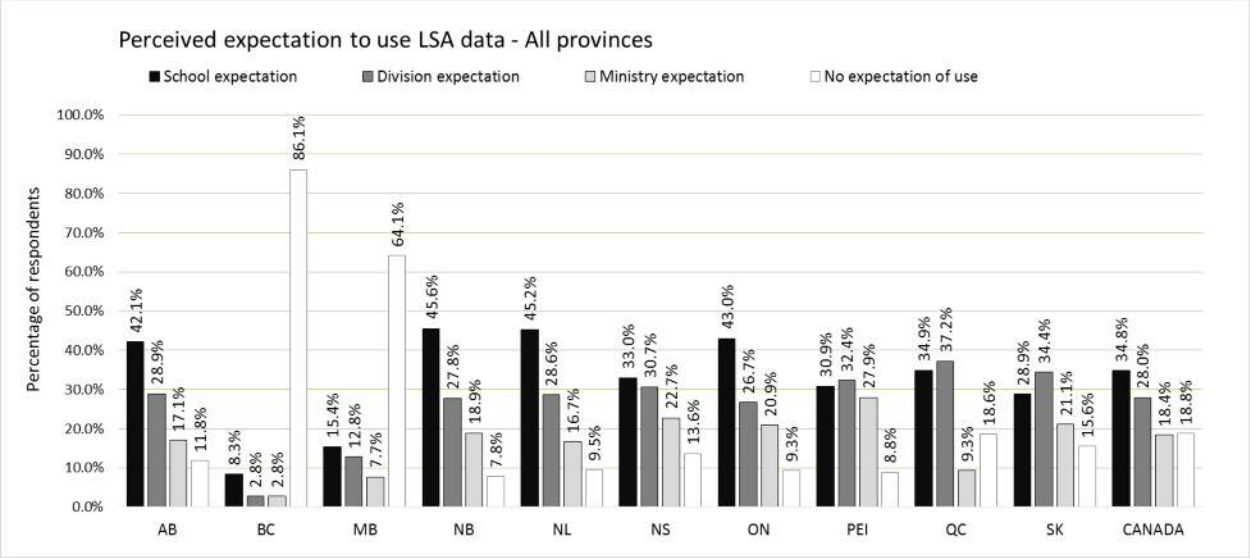

Figure A4: Provincial / national data on the perceived expectation to use LSA data.
The national average is relatively in line with most provincial numbers, but the low n for provincial survey data makes detailed analysis problematic.

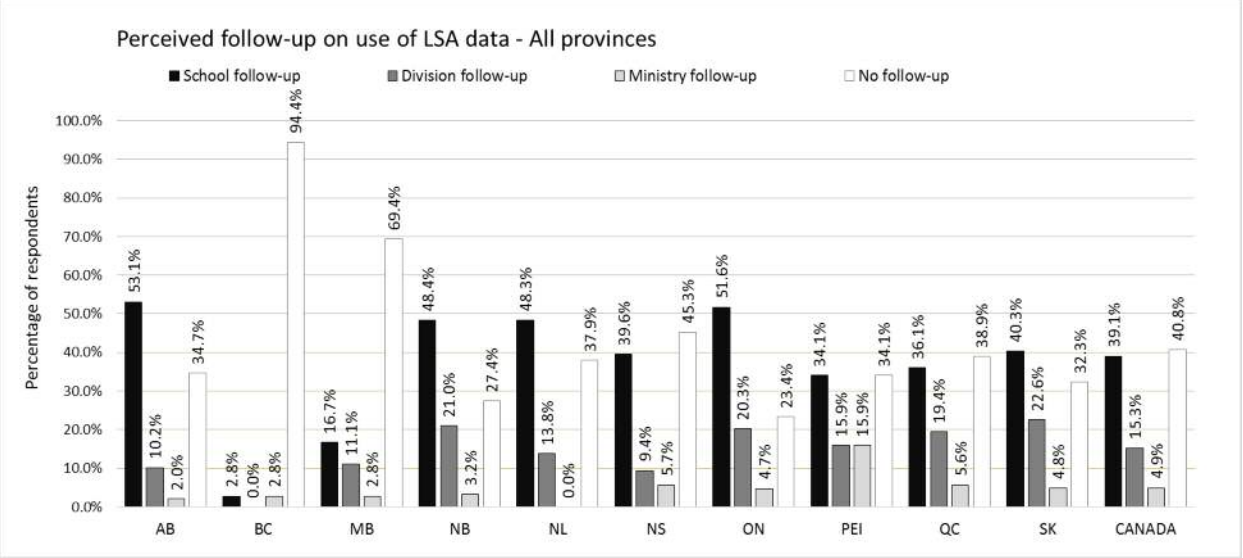

Figure A5: Respondents rated following up on instructional change.
National averages align well with most provincial responses.

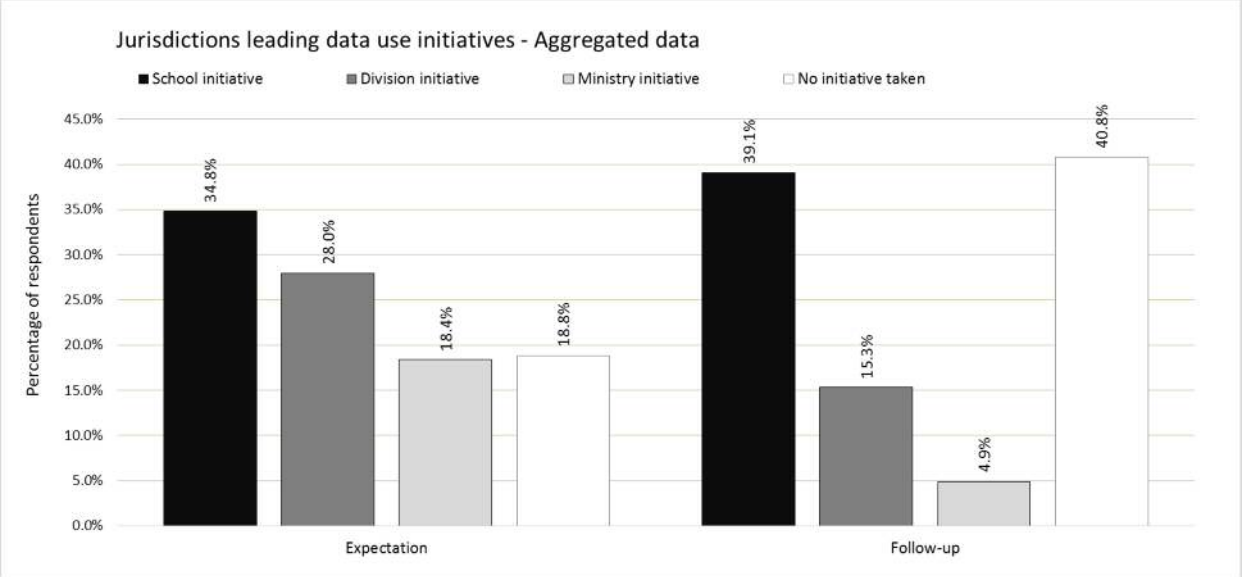

Figure A6: Breaking down expectations and follow up on data use by jurisdictions.
National data are shown here to illuminate from which jurisdictional level responding teachers reported expectations and follow up.

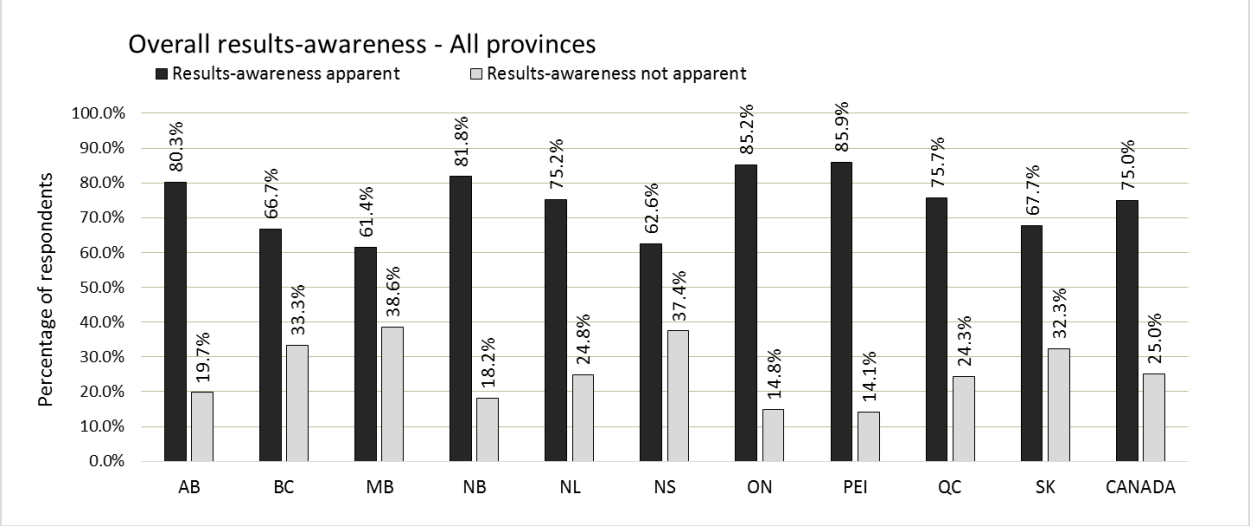

Figure A7: Averaging results-awareness scores across class, school and divisional levels. *Self-reported awareness is shown here, while awareness being not apparent came from respondents who did not know or receive LSA data.*

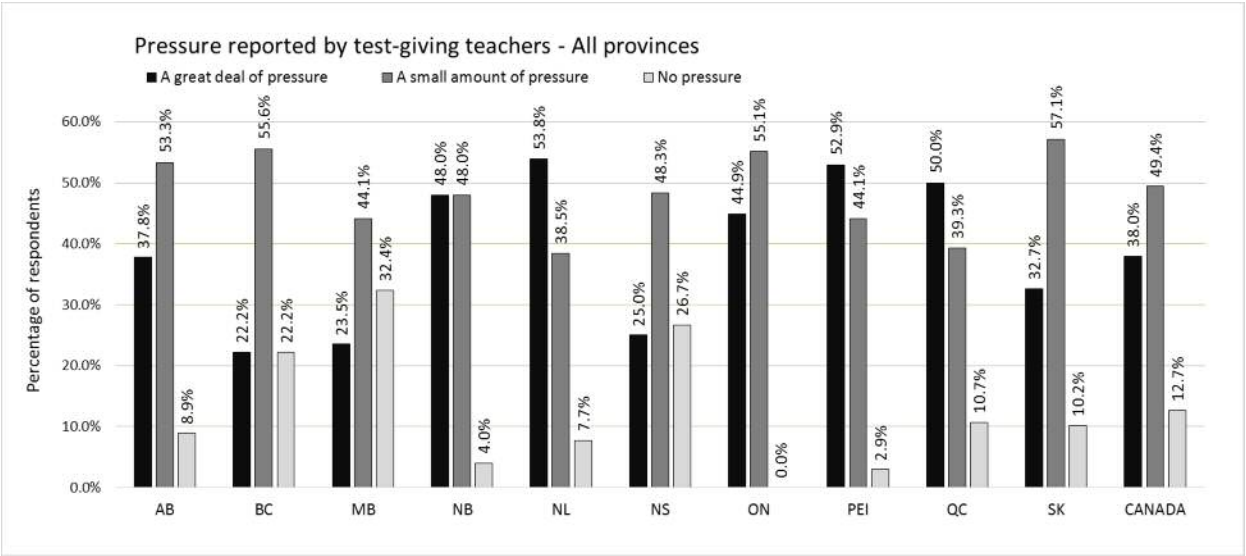

Figure A8: Teachers reported how much pressure they feel in relation to LSA testing.

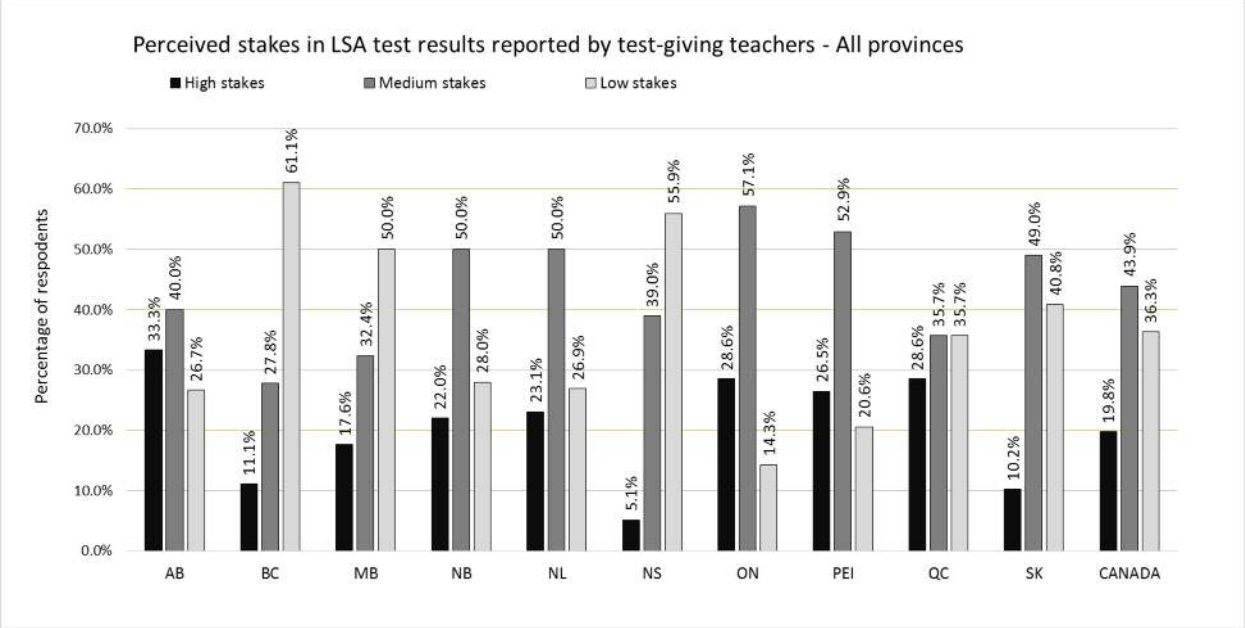

Figure A9: Teachers rated the level of stakes they think are applied to teachers by LSAs.

Appendix 2

The Survey Instrument

This survey was written for Survey Monkey which allowed it to be emailed to teachers and results compiled electronically. The square-bracketed values beside the survey responses indicate values assigned to responses which may help inform the analysis of regressions. Where they do not appear, they were not assigned or used in this fashion. Positive values were applied to factors thought to increase teacher use of LSA data. Lesser and negative values were applied to factors thought to impede or decrease the use of these data.

Both the survey and the interview data gathering covered more ground than was discussed in this paper. Other topics are found in papers that have been published or that should be forthcoming. This paper uses survey data from only the third (reactivity) and fifth (incentives) sections.

Background Information

1. Consent to terms of survey: Yes; No
2. Age: 18-24; 25-34; 35-44; 45-54; 55-64; 65+
3. Gender: Male; Female
4. Grades taught: Kindergarten or pre-K; Elementary (grades 1-5); Middles years (grades 6-8); High school (grades 9-12)
5. Teaching experience: 0-4 years; 5-9 years; 10-14 years; 15-19 years; 20-24 years; 25+
6. School setting: Urban setting; Suburban; Rural; Northern or remote
7. Staff size: Less than 15; 15-24; 24-34; 35-44; 45+
8. Class size: 1-10 students; 11-15; 16-20; 21-25; 26-30; 31+
9. Qualifications: College or high school; University but no BEd; University and BEd; Graduate studies but not completed Masters; Graduate studies with completed Masters or more
10. Subject area credentials: Indicate a major and/or a minor in – English; Mathematics; Science; Social Sciences; Other
11. School division
12. Voluntary email address (for selection of future interview respondents)
13. I give provincial tests in my classroom: Yes; No

Test Design and Results Data

14. Which tests does your class(es) take? English; Mathematics; Science; Social Studies; Other
15. Results are returned: The same school year [1]; The next school year [-0.5]; The results are not returned to me [-1]; I'm not sure [-1]
16. Results compare: Divisions with other divisions; Schools with other schools; Teachers with other teachers; Students with other students (for all 'Yes' responses [1]; all 'No' responses [-0.5]; all 'I'm not sure' responses [-1]; all 'I do not see the results' responses [-1])
17. Results are returned by: Department heads [1]; Administration [1]; Divisional personnel [1]; Ministry personnel [1]; I do not see the results [-1]; I'm not sure [-1]; Other (written response)
18. Are the results easy to understand? Yes, they are easy to understand as presented [1]; I have an incomplete understanding of the results as presented [-0.5]; I do not understand them as presented [-1]; I do not see the results [-1]
19. Can you act directly to use these result to inform your instruction? Yes, we can act directly [1]; Some interpretation is needed before we can act [-0.5]; We cannot act directly because teachers are responsible for analysis [-1]; We cannot act because the results are poorly or incompletely presented [-1]

20. Which kind of test items are used too much or too rarely: Selected response; Short constructed response; Longer constructed response (all 'Used too much' responses [-1]; all 'Used too rarely' responses [-1]; all 'Current use is appropriate' responses [1].

Using the Data (Reactivity responses, as reported in Author 2016b)

21. Have you ever been involved in writing or marking provincial assessments? Yes; No
22. Changes in instruction 1: I have looked for PD to improve my instructional strategies; I have requested additional resources related to testing; I have worked with other teachers to make sense of the data; I cover a wider range of topics in the curriculum; I hold group study session or provide extra help after school (all 'Not at all' responses [0]; all 'Somewhat' responses [0.5]; all 'A great deal' responses [1])
23. Changes in instruction 2: I cover material I know will be on the test very thoroughly; I focus more on test-taking strategies like 'the process of elimination'; I use the format of the test to give similar types of practice questions; I focus more on subjects that have provincial tests; I review old exams (all 'Not at all' responses [0]; all 'Somewhat' responses [-0.5]; all 'A great deal' responses [-1]).

Supports (as reported in Author 2017b)

24. Do you share results with following year's teachers: Never [-1]; Sometimes [0]; Always [1]
25. Are results from last year's teachers shared with you? Never [-1]; Sometimes [0]; Always [1]
26. Supports provided: PD; PLCs; Assessment teams; Administrative support; Printed or online guides; Coaching and/or mentoring; Other (written response) – (all supports as provided by any one of school, division, or ministry scored [1])
27. Helpfulness of supports: School supports; Division supports; Ministry supports (all 'Very helpful' responses [1]; all 'Helpful' responses [0.5]; all 'Not helpful' responses [-0.5]; all 'Not provided' responses [-1].

Incentives (reported in this paper)

28. Expectation to use data from: School administration [1]; Division [1]; Ministry [1]; No expectation of use [-1]
29. Follow up on use from: School administration [1]; Division [1]; Ministry [1]; There is no follow up on use [-1]; Other (written response)
30. Perceived pressure: None [0]; A small amount [0.5]; A great deal [1]
31. Results recollection/awareness: Class results; School results; Division results (all responses showing recollection [1]; all responses showing no recollection [-1]). Responses listed for each were: Better than average; About the same as average; Worse than average; I do not recall; These results are not provided.
32. Perceived stakes (for teachers and schools): High [1]; Medium [0.5]; Low [0].

Attitudes about Testing (reported in Author 2016a)
All responses in the following sections are scored -1 for 'Disagree', 1 for 'Agree', and 0 for 'Neither agree nor disagree.'

33. Provincial testing:	a. Is a good way to evaluate a school
(School Accountability)	b. Is an accurate indicator of a school's quality
34. Provincial testing:	a. Determines if students meet qualification requirements
(Student Accountability)	b. Makes parents better aware of student growth

	c. Selects students for education / employment opportunities
35. Provincial testing:	a. Identifies student strengths and weaknesses
(School Improvement)	b. Helps students improve their learning
	c. Is integrated with teaching practice
	d. Allows different students to get different instruction
	e. Changes the way teachers teach
36. Provincial testing:	a. Interferes with appropriate teaching
(Negative test attitudes)	b. Data only get used when stakes are high
	c. Has little impact on teaching practices
	d. Results are filed and ignored
	e. Is an imprecise process

37. Appropriate uses: Assign or re-assign students to classes; Identify learning needs of students who are struggling; Discuss student progress or instructional strategies with other educators; Form small groups of students for targeted instruction; Discuss data with a parent; Discuss data with a student; Choose which parents to contact; Meet a specialist about data – e.g. instructional coach (all 'Appropriate' responses [1]; all 'Not appropriate responses [-1])

About the Author

Derek T. Copp
Good Spirit School Division
derek.t.copp@gmail.com
Derek Copp has a PhD from Maastricht University's Graduate School of Governance. He is employed as a teacher and administrator at Esterhazy High School in the Good Spirit School Division (Saskatchewan, Canada). His research work is focused on the practical results of assessment policy, data-informed decision making, First Nations and rural education, and education policy generally.

education policy analysis archives

Volume 25 Number 115 November 20, 2017 ISSN 1068-2341