Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

Copp, Derek T.
The Impact of Teacher Attitudes and Beliefs about Large-Scale Assessment on the Use of Provincial Data for Instructional Change
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 24, 2016, pp. 1-25
Arizona State University
Arizona, Estados Unidos

education policy analysis archives

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

Arizona State University

Volume 24 Number 109 October 24, 2016 ISSN 1068-2341

The Impact of Teacher Attitudes and Beliefs about Large-Scale Assessment on the Use of Provincial Data for Instructional Change

Derek T. Copp
Good Spirit School Division
Canada

Citation: Copp, D. T. (2016). 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

Abstract: In the quest to improve measured educational outcomes national governments across the The Organisation for Economic Co-operation and Development (OECD) and beyond have instituted large-scale assessment (LSA) policies in their public schools. Controversy almost universally follows the implementation of such testing, related to such topics as: a) the uncertain quality of the tests themselves as psychometrics measures; b) the uses to which the data can and should be put; c) the unintended consequences of test-preparation activities and resulting score inflation; and d) the effects of high-stakes tests on students. Debates of this nature naturally involve and impact the attitudes and opinions of teachers related to their collection and use of these data. This paper examines the impact of these attitudes using both the qualitative and quantitative data from a large-scale research study on Canadian provincial assessment. Data were collected from nation-wide teacher surveys as well as interviews with teachers, administrators and district-level staff. Results show that teacher attitudes about these assessments are strongly correlated to classroom-level instructional change. Three attitudinal factors have significant effects on teaching (to) the provincial curricula, yet none significantly affects the use of less constructive instructional strategies also known as 'teaching to the test.' Specifically, the belief that large-scale assessment data have

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

Manuscript received: 4/28/2016
Revisions received: 7/27/2016
Accepted: 7/27/2016

more appropriate uses and the belief that these data could lead to school improvement were significant factors in facilitating change. The implications of these findings are profound in that large-scale assessment policy cannot succeed even by its own standards without more buy in from teaching professionals.
Keywords: Teacher attitudes, teaching to the test, large-scale assessment, reactivity, Canadian provincial assessment, data-informed decision making

El impacto de las actitudes y creencias de los profesores sobre la evaluación a gran escala y el uso de los datos provinciales de cambios en la instrucción

Resumen: Para mejorar los resultados educativos Organización de los gobiernos nacionales para la Cooperación y el Desarrollo Económico (OCDE) estableció la evaluación a gran escala de las políticas (LSA) en sus escuelas públicas. La controversia por lo general sigue estas pruebas debido a: a) la calidad incierta de las pruebas psicométricas como medidas; b) los usos previstos y deseados de los datos; c) las consecuencias no deseadas de las actividades de preparación de la prueba y la inflación resultante puntuación; y d) los efectos de las pruebas de gran importancia en los estudiantes. Estos debates afectan las actitudes y opiniones de los profesores sobre la recopilación y uso de estos datos. Este artículo examina el impacto de estas actitudes que utilizan tanto la cualitativa y cuantitativa de datos a partir de un estudio de investigación a gran escala sobre la evaluación de cada provincia del Canadá. Los resultados muestran que las actitudes de los maestros acerca de estas evaluaciones están fuertemente correlacionados con el cambios de instrucción en el aula. En concreto, la creencia de que los datos de evaluación tienen usos más apropiados y la creencia de que estos datos podrían conducir a la mejora de la escuela fueron factores importantes en la facilitación del cambio. Las implicaciones de estos hallazgos son profundas, en que la política de evaluación no puede tener éxito incluso por sus propias normas, sin más apoyo de los profesionales de la enseñanza.
Palabras-clave: actitudes de los profesores, la evaluación a gran escala, la reactividad, la evaluación de cada provincia del Canadá, la toma de decisiones basada en datos informados

O impacto das atitudes dos professores e crenças sobre avaliação em larga escala ea utilização de dados provinciais para mudanças na instrução

Resumo: Para melhorar os resultados educacionais governos nacionais da Organização para a Cooperação e Desenvolvimento Económico (OCDE) instituiu políticas de avaliação em larga escala (LSA) em suas escolas públicas. A controvérsia geralmente segue esses testes, porque: a) a qualidade incerta de testes psicométricos como medidas; b) as utilizações previstas e dados desejados; c) as consequências não intencionais das atividades de preparação para o teste e marcar a inflação resultante; e d) os efeitos de provas de grande importância para os alunos. Estes debates afetam as atitudes e opiniões dos professores sobre a coleta e uso desses dados. Este artigo examina o impacto dessas atitudes usando dados qualitativos e quantitativos de um estudo de investigação em grande escala sobre a avaliação de cada província do Canadá. Os resultados mostram que as atitudes dos professores sobre estas avaliações são fortemente correlacionada com a alteração do nível de instrução em sala de aula. Especificamente, a crença de que os dados de avaliação em larga escala são os usos mais apropriados e a crença de que esses dados poderiam levar a melhoria da escola foram fatores importantes na facilitação da mudança. As implicações destes resultados são profundas em que a política de avaliação não pode ter sucesso até mesmo por seus próprios padrões, sem mais apoio dos profissionais do ensino.
Palavras-chave: atitudes do professor, ensinando para o teste, a avaliação em larga escala, a reactividade, Canadá, a tomada de decisão informada por dados

Introduction

In the quest to improve measured educational outcomes, national governments across the OECD and beyond have instituted large-scale assessment (LSA) policies in their public schools. In all of Canada's ten provinces (which have independent jurisdiction over educational policy) LSA has also become a standard tool used to measure educational effectiveness and to make teachers and schools accountable for academic performance. Public and sector-specific debate almost universally follow the implementation of such testing. Researchers and the public alike question: the uncertain quality of the tests themselves as psychometrics measures; the uses to which the data can and should be put; and, the unintended consequences of test-preparation activities and resulting score inflation. High-stakes tests also substantially increase the pressure on students to perform. All of these factors have a role in shaping teacher attitudes and opinions on the collection and use of these data. It is this factor, teacher attitudes and beliefs about LSA, which is the focus of this paper.

The author completed a research study of how Canadian teachers use LSA data to improve their instructional practices (Copp, 2015). That larger study included inquiries into several different factors thought to have an impact upon the use of LSA data in classrooms. Some of these factors showed no significant links to the use of data, while others promoted both the defensible and the questionable types of instructional strategies. There were two examined groups of variables that did have significant statistical relationships to the use of data and also steered the instructional improvements towards those that are widely agreed to be both ethical and broaden the number and variety of outcomes presented to students. In short: a) attitudes/beliefs variables; and b) supports variables were statistically linked with teachers teaching (to) the curriculum, rather than teaching to any given test. The former will be examined here.

What follows is laid out in the following way: a) the possible instructional effects of LSA are explored; b) a literature review of studies related to teacher attitudes/beliefs is presented; c) the reactivity framework for the analysis is identified; d) the research questions and methodologies are discussed; e) the results of the data collection survey and interviews are given; and finally, f) the conclusion will look at the implications of this and future related research in this field.

Instructional Effects of Large-Scale Assessment

It has long been noted in the assessment literature that the mandatory administration of large-scale assessment has a range of effects, some of which are considered to be more educationally defensible than others. Cizek (2001) made note of several serious negative consequences of external testing including: (a) the reduction of time for regular instruction; (b) the side-lining of those topics/subjects not covered in LSAs; (c) the over-use of assessment strategies rooted in specific test designs; (d) poor staff morale; and (e) the use (or threat of use) of LSAs to punish students. Other 'unintended consequences' of LSA are quite extensively covered in the literature: (a) teaching to the test; (b) coaching; (c) curriculum narrowing; and (d) outright cheating (Darling-Hammond & Rustique-Forrester, 2005; Fehr, 2008; Luna & Turner, 2001; Simner, 2000; Volante, 2007; Volante & Cherubini, 2007). Coburn and Turner (2011) illuminate some of the factors, constraints and processes that make the use of data more or less possible and productive (such as accountability policies, school-based data use routines, or adding stakes to results).

Some of these practices are no doubt utilized with the expressed intention to improve LSA scores. Some, though, may well be used as a response to the data when no other path is obvious or made clear. Koretz and Jennings (2010) try to make clear the line between appropriate and inappropriate test preparation practices (a key point, so quoted at length):

> Preparation that leads to improved mastery of the domain is appropriate. This will necessarily produce score gains that show at least a modicum of generalization to performance in other contexts, including but not limited to other tests. Preparation that generates gains largely or entirely specific to the given test is inappropriate because it generates inflation rather than meaningful gains. (p. 18)

The figure below, adapted from Haladyna, Nolen, & Haas (1991, p. 4), specifies instructional strategies and rates their ethicality, and their effect on student outcomes (reprinted and supplemented with the authors'permission).

Test preparation activity	Degree of ethicality	Effect on student outcomes
Training in test-taking skills	Ethical	Narrowing
Checking answer sheets for proper completion	Ethical	None
Increasing motivation by appealing to students, parents	Ethical	None
Developing a curriculum based on the content of the test	Unethical	Narrowing
Preparing objectives based on items on the test	Unethical	Narrowing
Presenting assessment items similar to those on the test	Unethical	Narrowing
Dismissing low-achieving students on test day	Highly unethical	None
Presenting items verbatim from the test to be given	Highly unethical	Narrowing

Figure 1: Ethical and unethical educational practices and their effect on the number and variety of student outcomes

It is important to note that not all instructional change is either test-focused and/or unethical. Test results can be a guide to instruction, and they can ensure that the prescribed curriculum is well covered. Volante (2004) considers broad curriculum coverage (used as a means to improve achievement scores) suitable preparation. When LSA data are used as a means of making appropriate instructional adjustments for improvement, they can be a key element of professional improvement (Loughran, 2002). These data can and often do guide instruction in positive ways.

Good teaching practices such as monitoring, questioning, and providing feedback are also ways that teachers can stimulate achievement gains (Gibson & Dembo, 1984). Volante (2005) lists some of the many ways individual teachers can effectively use these data: (a) to assess their own pedagogy; (b) to indicate what might prove to be relevant professional development (PD); and (c) to give evidence of high-quality school programs and the allocation of resources. Darling-Hammond and Rustique-Forrester (2005) also note several positive changes that LSA results can trigger for schools: (a) greater awareness of curriculum standards by school leaders; (b) greater attention given to students who need support to improve results; and (c) to indicate what domains are most important to teach and learn by setting clear, specific goals and providing feedback.

The Role of Teacher Attitudes in Professional Practices

There are many policy goals set out for Canadian LSAs (including but not limited to): a) that the tests are written across various subjects and grades to gauge progress; b) that the results data are distributed to schools and teachers; and c) that teachers adapt and improve their instruction based on the results data. Yet the expectation of instructional change depends upon solid, sustained policy implementation, which itself depends in no small measure upon the willingness of teachers to accept these data as a sound basis for change. The implementation of LSA policies depends upon several mediating factors, including the teacher attitudes variables which are considered in this paper. The effects of these variables are considered in several recent research studies.

A recent study by Brown, Lake and Matters (2011) noted that teacher attitudes and preconceptions are important to implementation choices. This study was built upon Brown's (2004) prior work which used survey questions to allow teachers to self-report their opinions of LSA. The authors noted that data-use practices might be choices based not on policy factors, but on principles:

> The antipathy for holding both schools and individuals accountable through assessment may not be so much a desire to escape responsibility, but rather a rational rejection of poor-quality assessment systems that have unjust stakes or consequences for schools and/or learners. (Brown, Lake, & Matters, 2011. p. 218)

Remesal (2011) stated that teachers' beliefs were often of 'mixed conception' and that, despite research that indicates the importance of beliefs, LSA policies tended to zero in on 'assessment competence' or 'assessment literacy' rather than addressing attitudes. A quantitative study by Segers and Tillema (2011) showed a wide-range of attitudes regarding assessment across a sample of Dutch educators, some of which were not supportive of an 'assessment for learning' (AfL) framework in schools.

Kitaishvili (2014) noted that using professional development to enhance not only skills but to improve attitudes about learning and assessment was on way to deal with resistance to change:

> Teacher professional development first of all should be focused on changing teachers' attitudes and conceptions of assessment to ensure effective adoption of a desired practice. Teachers need to make many changes to their thinking as well as continuously reflect on their changes and achievement; they need intensive practice to gain skills and competencies important for using various assessment approaches that require cognitive complexity from their students. (p. 173)

Cobb & Jackson (2015) also stress that the kind of PD provided is just an important as its availability. Another interesting examination of teacher attitudes was completed by McMillan and Nash (2000) which identified several mediating factors between policy expectations and the fidelity of teacher implementation. LSA is very commonly defended as the primary educational accountability tool (see, for example: Tobin, Lietz, Nugroho, Vivekanandan & Nyamkhuu, 2015). The recent work of Datnow and Hubbard (2015) cited Park (unpublished, 2008) in the assertion that one of the most prominent beliefs teachers hold about LSA is that it is intended to address political accountability, not educational needs. The authors go on to state that teachers find numerous technical test design flaws in the tests themselves (also noted in: Schifter, Natarajan, Ketelhut, & Kirchgessner, 2014) which goes some way in explaining why these assessments are considered more useful at the policy level than in the classroom. The use of tools unsuited to their assigned purpose is noted by Marion & Leather (2015) as one reason that organizational change (i.e. improved instruction) is not apparent in systems with accountability-based LSAs.

Reactivity Framework

This study is based on the theoretical concept of reactivity. Campbell (1957) examined reactivity as one of many possible design flaws in experimental methodology. Commonly known as 'the Campbell effect,' reactivity is the quantifiable change in behaviour when people know they are being observed or evaluated. These reactions, whether conscious or unconscious, have an impact on the objectivity of the measurement. It is not a surprise that teachers are reactive to external assessment, especially when sanctions or rewards are written into policy. Such policy-based incentives are supported by economists like Hanushek and Raymond (2002):

> The rewards and sanctions that many states have built into their accountability systems create the motivation for schools to change behavior. (p. 16-17)

Examining how teachers react, whether it is in ways that improve scores on tests at the expense of the non-tested content, teaching to the test, or in ways that improve overall student learning was one of the primary purposes of this research study.

Provincial assessment policies tend to trust Hanushek and Raymond's assertion (above) as well as the common assumption in LSA score analysis: higher test scores means that better learning has occurred. The effectiveness of educators is thus measurable and can be quantified based on scores from provincial tests. With a clear focus on this metric, teachers then have a principled choice to make between reactivity options. The distinction between types of reactivity made in this paper is also based on principles, namely those of the local teaching authority in the author's jurisdiction – the Saskatchewan Teachers' Federation (2015). This Code of Professional Competence spells out clearly what kinds of instructional practices are expected from competent educators. It is also in keeping with other provincial codes of professional conduct and broadly aligned with what are the policy directives of all provinces for their LSAs (Copp, 2016). The ten survey-listed instructional strategies are consciously divided based on these principles and the knowledge that commonly utilized instructional strategies can be test-focused or curriculum-focused, but rarely qualify as both (see Koretz, above) as a result of the practical constraints on LSA design and scoring.

Teaching (to) the curriculum includes practices which are thought to be both ethical and broaden the number and variety of outcomes presented to students (see appendix, table 4 for reactivity scores). This approach is less likely to result in higher scores on any single test, but it would provide skills to improve achievement in more situations since it avoids the potential pitfalls of teaching to a specific evaluation instrument (Popham, 2001). Strategies in this category qualify under the terms of the STF Code of Professional Competence in all regards.

Teaching to the test encompasses those educational practices which are thought to be are either unethical or reduce the number or variety of outcomes presented to students. These methods are certainly the most direct way to improve scores on a specific test, but even practices in this category which might be called ethical do not have the transferability, the increased 'leverage,' upon which high road practices are premised (Au, 2007; Jacob, 2002). These strategies do not, when used by default, meet the terms of the same STF code of conduct.

It has been widely noted in the literature and it is pivotal to this distinction between grouped strategies that depending on the test-preparations practices employed, improved scores on provincial LSA tests might indicate something about students' learning, but they may not always mean that more or better learning has occurred (Linn, 1998; Koretz, 2002). It is possible that higher scores, counter-intuitively, might indicate less or narrowed learning. Even though LSAs are designed to provide a clear accounting of educational achievement, they may not effectively serve this primary purpose. Amrein and Berliner (2002) note the role of Campbell's law in clear terms:

> The more important that any quantitative social indicator becomes in social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor… [Thus] attaching serious personal and educational consequences to performance on tests for schools, administrators, teachers, and students, may have distorting and corrupting effects. (p.5)

This unique reactivity model was designed to account for these factors and to determine the level of teacher responsiveness. The type of reactivity was also calculated as the dependent variable. Secondary lines of inquiry not addressed in this paper included teacher supports, policy incentives, test design and data, as well as background factor variables such as age and experience. All data were self-reported by survey respondents, not observed first-hand (appendix, table 4 has provincial average scores).

Figure 2 shows the specific survey prompts used to determine types of reactivity employed in classrooms. The left is designated teaching (to) the curriculum, and the right side of the chart is named teaching to the test. Respondents were asked to consider how their instruction may have changed in classes that write provincial assessments, and state how much they had used these strategies. Options for responses were 'not at all', 'somewhat', or 'a great deal.' It should be noted that no instructional practice listed here is said to be more or less effective in any specific case. Teachers use their professional judgement every day to make instructional decisions such as those regarding the use of data. What is clear is that a regular and consistent emphasis on teaching to the test strategies does narrow the curriculum made available to students and also limits the means by which they can demonstrate their understanding of curriculum outcomes (as in figure 1).

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 2: Reactivity prompts from research survey

Note: The full survey covered several lines of inquiry as well as reactivity which was the dependent variable of the original study (see appendix). Teacher attitudes (related to reactivity effects) is the line of inquiry covered in this paper.

Research Questions and Methodology

With considerations in mind about the different testing models, the varied ways that teachers might react to the results, and the importance of prior opinions and beliefs in mind, this paper sets out to answer the following research questions:

1. Are teachers' attitudes regarding large-scale assessment correlated to their use of LSA results data in their classrooms?
2. Is classroom data-based decision making correlated with the more educationally defensible practices or towards teaching to the test?
3. Which attitudes variables are correlated with the instructional use of LSA data?

Both survey (quantitative) and interview (qualitative) data are used to inform the discussion. Mixed methods were thought to fit these data best in order to both identify statistical correlations and then to draw out more clarifying details from interview respondents (McMillan & Schumacher, 2010). The mixed methods model employed was a sequential explanatory design in which the primary data source was a survey instrument and results were collected for analysis. The second phase of data collection consisted of interviews, which were coded in terms of their explanatory value and to augment the findings of the quantitative analyses (McMillan & Schumacher, 2010).

Surveys

Surveys were emailed to participating school divisions in all ten Canadian provinces and to schools at all grade levels K through 12. A review of the literature uncovered useful field-tested questions from research studies: (a) questions about reactivity strategies were adapted from Skwarchuk (2004), and Hamilton and Berends (2006); (b) teacher attitude questions were adapted from Brown (2004); (c) questions about the use of assessment data were adapted from Wayman, Cho, Jimerson and Spikes (2012); and, (d) questions regarding supports and professional development were adapted from Boyle, Lamprianou and Boyle (2005).

This study used a cross-sectional design since surveys were sent in the 2013-2014 school year. The sampling unit was individual teachers, Canadian public school teachers were the target population, and clusters for the (probability) sampling were school divisions. The target population was geographically divided into non-overlapping groups (divisions) which resulted in cluster sampling (see figure 3). All teachers from participating schools (whether or not they administered LSAs) were asked to take part. The selection of participants was in name random, however it was greatly affected by the reality that many school divisions chose not to take part. Thus clusters were as much a voluntary sample as a random one.

The target population included several strata: (a) teachers from urban and rural areas; (b) teachers in large and small schools; (c) teachers at different grade levels; (d) teachers of different ages and sexes; (e) teachers of different subjects; (f) teachers with varied amounts of experience; (g) teachers with varied qualification levels; and (h) teachers with large and small classes. Single province samples were too small for in-depth analysis, but the larger *n* of the nation-wide data meant that all strata were represented. There is a distinct lack of demographic data available for teachers at the national and provincial levels. Statistics Canada collects 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 (for an admittedly limited number of comparators) lends some weight to the ability of these results to generalized to the wider Canadian teaching population (McMillan & Schumacher, 2010). It bears noting that the response rate was much higher for teachers who give LSAs who had the first-hand knowledge needed to respond (considering that approximately one in four teachers do not administer themselves these tests).

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%

Figure 3: Survey response rates from all provincial jurisdictions.

Note: For data analysis, partially-completed surveys were removed from these figures. It is notable that nearly half of respondents themselves administered LSAs.

Schools and school divisions receive many requests for research studies, so the division and school level administrators were often hesitant to approve the study. Each province is made up of between one and 31 school divisions, and applications for research were sent to nearly one hundred of these across Canada. Successful applications numbered 27.3% (27/99). This assent gave the researcher to access school administrators. About 25.1% of school administrators contacted chose to participate (181/720). Note that none of the non-responses above should be counted against the response rate since these teachers were not given the option to respond. When emailed the link from their principal and given the choice to participate, 23.1% (453/1963) responded to the survey of 1963 who were counted as those who administer LSAs. Data were also collected from teachers who do not administer LSAs, but these data were not used in this particular study.

While all teachers in a given school had the opportunity to respond, the research questions were aimed more directly at teachers who administer LSAs in their classrooms, and the minimum number of respondents in this group from each province was set at 30. Small sample sizes did not support in-depth analysis of provincial data, but national numbers were sufficient for these procedures. While the overall response rate is low overall (Nardi, 2006), Olsen (2006) has questioned whether non-response itself casts serious doubt on any reasonably responded-to instrument.

The survey instrument was checked for internal validity using two distinct statistical tests. As each line of inquiry (groupings of independent variables) was targeted to different features of tests and data use, Spearman's rank order correlation tests were employed to determine what relationships exist between them, which relationships complement each other (positive correlations), and also which relationships are at odds (negative correlations). This was one means to verify that the binary values applied to survey responses were appropriate, and also to determine if there were covariant variables in the study. All lines of inquiry showed several significant and positive correlations but none went so far as to be distorted by multicollinearity.

The Spearman's values for questions related to teacher attitudes are shown in table 1. Significant positive correlations are apparent (excepting negative test attitudes which was counted in

negative integers) but not so high as to show multicollinearity (the student and school accountability variables show the strongest relationship here, that of 0.480, significant at the $p>0.01$ level).

In order to address the internal consistency of the dependent variable (in its two distinct forms), Cronbach's alpha was utilized. It is often used in data comprised of aggregated scales (both teaching (to) the curriculum and teaching to the test scores are aggregated from several survey responses) to determine the degree to which the survey items appear to measure the underlying construct (Santos, 1999). Cronbach's alpha is this sort of index of reliability, and showed reasonable levels of internal consistency for both measures. Alpha scores of 0.7 or more indicate adequate congruence of the items in the aggregated score. Values less than 0.7 are less certain in terms of their adherence to the underlying construct being measured. The alpha for teaching (to) the curriculum items was 0. 62 and for teaching to the test items was 0.76. Scores for all items in both scales were quite consistent so that eliminating any one item did not generate a significantly higher alpha. These values are subject to interpretation which can be left to the reader except to note that having a small number of values in aggregated scales may result lower alpha scores (Tavakol & Dennick, 2011). Both of these scales consisted of five items.

Table 1
Matrix of from Spearman's rank order correlation test with attitudes variables

1. School accountability	1.000				
2. Student accountability	0.480**	1.000			
3. School improvement	0.358**	0.519**	1.000		
4. Negative test attitudes	-0.286**	-0.285**	-0.456**	1.000	
5. Appropriate uses for data	0.160**	0.266**	0.287**	-0.217**	1.000

* p<0.05; ** p<0.01

The dependent variable (*Y*) in this study was teachers' use of LSA data, and the multiple regression equation, which follows the assumptions of the OLS for multiple regressions, is as follows (Stock & Watson, 2007):

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

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

Interviews

Using a semi-structured format, the interview guide was employed to follow up upon the same themes as the survey instrument. The intention was to triangulate the qualitative and quantitative data (Flick, 2006; Jick, 1979). While interviews were done to explore in more depth the lines of inquiry from the survey, it is also true that both commonalities and contradictions were apparent. These accurately indicate the wide range of opinions on the topics covered. The data set was seen as mutually complementary since qualitative data can help explain relationships noted from quantitative methods (Onwuegbuzie & Leech, 2005).

Subjects for interviews were purposively selected to fairly represent the range of choice in instructional strategies employed by teachers (Flick, 2006). Subjects were selected based on a stratified purposive sampling which examined the high and low ends of the reactivity score range as distinctive strata. There was also an element of multilevel sampling since in-school and division-level administrators were included in the second phase data collection (McMillan & Schumacher, 2010). Only classroom teachers had completed the survey (the reactivity scores from which guided their selection), but both in-school and division-level staff were included to gauge the congruity of their responses to those from front-line staff.

The sample size for interviews was quite small but included teachers from across Canada as well as school-based and divisional administrators. Interviews were conducted with 13 classroom teachers who administered LSA tests, 10 in-school administrators, and four division level staff. As with survey respondents, subjects were assured anonymity and confidentiality. They are identified in this paper only by subject and grade level taught (or a generic job description for non-teachers), province, and sex. Respondents from the small sample were not chosen to be representative of the survey population or to meet external validity standard set by McMillan and Schumacher, 2010. They were chosen purposively, as per Flick (2006) to provide extra detail and insight into the quantitative results. In this regard, interview data are not intended to be generalized, but quite the opposite.

Limitations

Being drawn from a larger research study, any flaws in the original come through in this focused examination. The data set from the survey was extensive but did lack variability in some locations as schools and/or school divisions chose not to take part. The data were collected under strict confidentiality and this left no room for school identifiers which might have made hierarchical linear analysis possible. It is also true that each line of inquiry followed may not be as extensive in the larger whole than if each topic had been studied individually but in greater depth. The interviews did provide valuable insights into the reactivity choices of teachers, but were drawn from a small purposively chosen sample, and thus have limited value beyond the insight they provide in conjunction with the survey data analysis. These limitations were and are solely the fault of the researcher and author.

Findings

Four Lines of Inquiry

The independent variables were identified initially from the literature on large-scale assessment included four main lines of inquiry operationalized from the survey for analysis purposes into numerical values: (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. Each line of inquiry consisted of several survey questions which were in turn given a numerical score based on the researcher's judgement. In general terms, if a variable was seen to promote the use of data, it was given a positive score (see appendix). Variables that discouraged or made more difficult the use of LSA data received lower or negative scores. Spearman's rank order correlation tests for each line of inquiry was completed as a check on the validity of these numerical scales, and the correlations proved the relative accuracy of the values placed. Correlations were regularly both significant and positive, but no distortion from multicollinearity was noted.

In order to perform an unbiased comparison of these lines of inquiry, all were standardized using an additive index algorithm and equated on a scale from 0-100 to ensure no discrepancies in the weight of the groupings. These additive index variables were used in multivariate regressions

alongside provincial dummies variables which were employed to help account for provincial differences in the data set. For each of the two defined varieties instructional change, the province with the average score closest to the national average for these effects was left as the control. This was intended to help arrive at values indicating which groups of variables had the most significant effects on different types of reactivity and how this varied across provincial jurisdictions. Table 2 shows the results of these computations.

Table 2
Regression table with the four lines of inquiry from survey data collection

	Teaching (to) the Curriculum		**Teaching to the Test**	
	Coefficients and (*t* scores) are shown			
Tests and data	0.0001	0.0023	0.0016	0.0039
	(0.03)	(0.56)	(0.34)	(0.84)
Supports	0.0143**	0.0124*	0.009	0.007
	(2.77)	(2.33)	(1.54)	(1.21)
Incentives	0.0142***	0.0141***	-0.0125**	-0.0109**
	(4.30)	(4.05)	(-3.24)	(-2.78)
Attitudes	0.0147**	0.0130**	-0.00182	-0.00401
	(3.22)	(2.77)	(-0.34)	(-0.77)
Provincial dummy variables	**AB**	0.109	**AB**	0.283
		(0.36)		(0.71)
	BC	-0.0738	**MB**	0.455
		(-0.19)		(1.08)
	NB	0.0482	**NB**	0.397
		(0.16)		(0.99)
	NL	-0.0876	**NL**	-0.128
		(-0.26)		(-0.29)
	NS	-0.298	**NS**	1.460***
		(-0.96)		(3.63)
	ON	-0.00437	**ON**	0.718
		(-0.01)		(1.74)
	PEI	-0.702*	**PEI**	0.0692
		(-2.33)		(0.18)
	QC	0.0185	**QC**	0.0520
		(0.06)		(0.12)
	SK	-0.629	**SK**	1.099*
		(-1.77)		(2.41)
Constant	0.475	0.649	-2.792***	-3.286***
	(1.40)	(1.59)	(-7.25)	(-7.47)
N	228	228	230	230
Adjusted R^2	0.199	0.231	0.033	0.148

*** p<0.05 **p<0.01 ***p<0.001**

What is revealed by these results is not unsupported by the literature. Koretz (2009) for example, discusses practical test design; Schildkamp and Kuiper (2010) examine various means to support data use; incentives are widely explored but have mixed results as in Finnigan and Gross (2007); and attitudes about data are the focus of both Brown (2004) and Brown, Lake and Matters (2011). An interesting relationship that appears is the fact that neither the design of tests nor the manner in which data are returned has any effect on reactivity of either variety. Across Canada the subjects and grades tested differ greatly (see appendix, figure 4), and seeing as how only high school level LSAs have significant summative weight, these differences might have been thought to lead to a significant finding. Considering how commonly cited test design and data return factors were by interview respondents, this was an even more surprising finding.

> The exam itself isn't meant to evaluate the entire curriculum and there is no way it can. There are items that are just above the scope of a multiple choice exam. You really can't ask a lot of open-ended questions on a multiple choice exam. . . And there are a lot of higher learning expectations in our program of study that require open-ended answers.
>
> *- AB, High school Science teacher, male*

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

Next, supports variables were found to have a significant correlation with teaching (to) the curriculum but no correlation with teaching to the test. When supports are provided, they appear to help steer teachers towards using the LSA data in more educationally defensible ways.

> Certainly at the school level and board level we have literacy consultants that work on a regular basis . . . that support our elementary teachers and middle school, too, but to more extent P-6 is well supported in terms of consultants going in, literacy support people going in to model and to coach the classroom teachers.
>
> - *NS, Division staff, female*

> To me, that is an administrator's job, to make sure that those teachers are comfortable, they're getting the time they need, they're getting the PD they need.
>
> - *MB, Elementary school principal, female*

Incentives have strong correlations to both positive and teaching to the test effects. It might be argued that 'incentives work' to promote instructional change, but these changes are not always in the intended or expected direction.

> They arbitrarily set our goal for this year to increase our writing average in our school by 3%, I think… So I said to them, if we succeed, then next year we'll have to raise it say 3% again so eventually we'll need to have a 100% success rate and a 105% average on each exam.
>
> - *QC, High school English teacher, male*

> 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*

It should be said that the Canadian context of stakes for LSAs is vastly different than that of other jurisdictions. Neither teachers nor schools are officially sanctioned as a result of poor results, and the only real stakes that are applied to the education sector come from public opinion following the publication of results. Students, of course, have a direct consequence for their final grades in some jurisdictions.

Finally, and most relevant to this paper, attitudes about LSAs and how the data might be used have a strong significant relationship with teaching (to) the curriculum, and none on teaching to the test effects. This factor is the most significant when it is taken into account that only teaching (to) the curriculum is a desired instructional outcome for teachers this being stated policy across all 10 Canadian education ministries (Copp, 2016). Attitudes are correlated only with teaching (to) the curriculum strategies, and the effects are more pronounced than for supports variables, which also point in this same direction.

It is worth noting that the second-step inclusion of provincial dummy variables has a much greater impact upon the adjusted R^2 figure for teaching to the test (they increase from 3% to 15% with the inclusion of the dummies) than does the same second-step procedure in respect to the teaching (to) the curriculum (R^2 values move only from 20% to 23%). This indicates that provincial variation has a lesser influence on the constructive instructional changes noted by teachers across all lines of inquiry. Provincial variation is, on the other hand, a major explanatory factor in accounting for the variance in results when we examine test-focused instructional strategies. Where the influence of these lines of inquiry appears nearly universal in terms of teaching (to) the curriculum, province-specific factors have much more impact on directed test-preparation activities.

Attitudes Variables

Having clarified the importance of teacher attitudes variables in decision-making about how to best use the data provided from provincial large-scale assessments, the next logical step is to examine the independent variables considered under the grouping 'attitudes' to see which have the most pronounced statistical influence (see table 3 below). Much along the lines of what has just been noted for the aggregated lines of inquiry regressions, the effect of the provincial dummy variables is much more pronounced on teaching to the test variables than on those linked to teaching (to) the curriculum.

There is a 14% increase in R^2 values after the second-step addition of dummies for teaching to the test variables, but a mere 3% swing when we look at teaching (to) the curriculum variables. At the risk of being repetitive, provincial variations in testing policy appear to be a major influence on teaching to the test variables, while these differences have a muted statistical effect on variables related to teaching (to) the curriculum factors. The results from Nova Scotia, Saskatchewan and Manitoba in particular show a great divergence from the national scores in that teaching to the test is much less common in respondents from these three provinces (i.e. the results show high levels of positive statistical significance).

The survey attitudes questions were strongly influenced by Brown's (2004) study involving New Zealand teachers' conception of LSAs. Four of the five factors included in the regressions from table 4 consisted of between two and five survey prompts which asked respondents to 'agree', 'neither agree nor disagree', or 'disagree' (a 3-point Likert scale). In addition to these factors, respondents were asked to rate which uses for the data they deemed appropriate or inappropriate from a list of 10 possible choices (adapted from Wayman, Cho, Jimerson and Spikes, 2012).

Table 3

The effects of attitudes variables are correlated to the use of both 'teaching (to) the curriculum' and 'teaching to the test' strategies.

	Teaching (to) the Curriculum		**Teaching to the Test**	
	Coefficients and (*t* scores) are shown			
School accountability	0.0643	0.101	0.110	0.0335
	(1.24)	(1.93)	(1.81)	(0.58)
Student accountability	-0.0724*	-0.0658	-0.0828*	-0.0744
	(-2.00)	(-1.78)	(-2.00)	(-1.87)
School improvement	0.0898***	0.0772**	-0.0243	-0.0327
	(3.54)	(2.95)	(-0.83)	(-1.15)
Negative test attitudes	-0.0825**	-0.0657*	0.0532	0.0164
	(-3.29)	(-2.58)	(1.85)	(0.60)
Appropriate uses for data	0.0441***	0.0393**	0.00169	-0.0100
	(3.55)	(3.18)	(0.12)	(-0.75)
Provincial dummy variables	**AB**	0.658*	**AB**	-0.107
		(2.54)		(-0.39)
	BC	-0.242	**MB**	0.790**
		(-0.87)		(2.67)
	NB	0.484	**NB**	0.191
		(1.91)		(0.69)
	NL	0.337	**NL**	-0.147
		(1.17)		(-0.47)
	NS	0.274	**NS**	1.083***
		(1.02)		(3.78)
	ON	0.442	**ON**	0.554
		(1.67)		(1.91)
	PEI	0.0410	**PEI**	-0.375
		(0.17)		(-1.41)
	QC	0.312	**QC**	0.0484
		(1.09)		(0.16)
	SK	-0.0568	**SK**	1.049***
		(-0.22)		(3.78)
Constant	2.484***	2.312***	-3.069***	-3.415***
	(28.42)	(11.52)	(-30.13)	(-15.78)
N	344	344	347	347
Adjusted R^2	0.163	0.194	0.027	0.170

*** p<0.05 **p<0.01 ***p<0.001**

The researcher-designated numerical value for individual responses were aggregated into total scores for each of the five independent variables. Provincial dummies variables were incorporated in order to account for provincial differences, and in each case, the designated control was the province with the average score closest to the national average to allow for both positive and negative correlations.

Teaching (to) the Curriculum:

The two variables that have equally strong statistical significance in terms of teaching (to) the curriculum are a belief that LSA data can be used for school improvement, and the belief that these

data can be employed for several appropriate uses for schools and teachers (see appendix, figure 4 for detailed data). These correlations are strong both before and after the addition of provincial dummy variables. It is clear that those respondents who had a firm belief in the potential use and value of LSA results were the most likely to use them by utilizing the instructional strategies thought to be best suited to meeting those goals.

> I would personally like to see it [more provincial testing]. I know as a school here, we collect all of the . . . mathematical exam information and test exams . . . and the same for literacy, we do it each quarter. We collect all the information on all the students, just to see how we are progressing. I think there is an accountability factor not only for the students, but also for the classroom teachers.
> - *NB, High school principal, male*
>
> From an accountability perspective, it certainly gives you a sense of how your kids are learning in your building compared to kids in other buildings, not just in your jurisdiction, but, you know, it compares apples to apples.
> - *ON, High school principal, male*

With a lesser significance level, it is also true that respondents who reported negative attitudes about tests (i.e. those who questioned their value or their validity) were less likely to employ teaching (to) the curriculum strategies.

> It is the day-to-day assessment that really determines whether the kids, you know, whether their achievement changes. It is not the provincial assessment. And so I think that is the, it is the double-edged sword if you want to say. So I would never say just because of the provincial assessments the kids' achievement changed.
> - *NS, Division staff, female*
>
> Like any test it is a one day, couple of hours, umm, some people don't do well in pen and paper… There is more than one way to assess an outcome and this seems to be only one way, that seems to be pretty limiting right there.
> - *PEI, K-9 school principal, female*

Alberta is the only province that deviates significantly from the national norm, yet the significance level of this relationship is low. As previously noted, these variables account for almost 20% of the variance in survey results, which is a reasonably high result for social sciences quantitative research.

Teaching to the Test

Switching over to the teaching to the test side of the table, there is only one independent variable that shows significant results at all, and in this case, only prior to the addition of provincial dummies. A belief that LSA data were an effective means of keeping students accountable led to a fleeting result indicating more use of teaching to the test strategies, and interestingly, a similarly fleeting result showing a diminished use of teaching (to) the curriculum strategies. While these mirror-image results do not stand up to the scrutiny of adding the provincial dummies, an interesting dynamic should be noted. Teaching (to) the curriculum strategies can be seen to involve more 'heavy lifting' by teachers, as they are generally time-intensive for educators. Teaching to the test strategies are much like picking the low-hanging fruit in that they do not necessarily involve more time or planning by teachers, just a different focus. Respondents who indicated a belief that students should themselves be held accountable by LSA results might then also be those teachers who are most willing to place the extra burden of time and effort for achievement score gains upon the shoulders

of students rather than themselves. The practice of teaching to the test itself highlights conflicting values from within the teaching community.

> I know there are teachers who teach to the test and cover graphing to make sure that the kids know about graphing because the graph question is worth four points. But I don't believe in teaching to the test. I like to cover the material as I think my kids are ready for it.
>
> - *BC, Elementary homeroom teacher, female*
>
> We took the time to say, okay, these are the types of questions you are going to get the types of what you will actually see, not content-wise. So, I don't look at that as teaching to the test, but you have to teach to the *style* of the test…So we took a lot of time to look at what, how are they going to ask the questions and, ahh, and what types of response will they be looking for.
>
> - *NB, High school principal, male*
>
> They had maybe five tests to do, which is a lot for 8 year olds… We spent most of the year prepping for the test, teaching to the test which is not really preference of teaching practice.
>
> - *AB, Elementary English teacher, female*

Three provinces deviate from the national norms in highly significant fashion. The reasons for these results might well be explained by noting that Nova Scotia and Saskatchewan do not have high stakes tests as part of their current policies. This would appear to have the result of constricting the apparent incentives to use teaching to the test practices. Manitoba does have high stakes tests for grade 12s, but the LSAs in other grades are evaluations unlike those done in any other province. These 'teacher checklists' are not pencil and paper tests students must complete, but periodic check-ins on curriculum outcomes with no direct relation to summative evaluations. These LSAs also seem to remove the motivations that normally accompany the use teaching to the test strategies.

Conclusions

This paper set out to examine one of the lines of inquiry followed in a nationwide survey of Canadian teachers in order to determine if respondent teachers' attitudes regarding large-scale assessment affected their willingness to use the results data from LSAs in an effort to improve their instruction. It was also important to try to establish what kind of instructional strategies were most closely linked to attitudes variables, and also which of several disaggregated variables had the greatest correlation to these effects.

What is clear from the data is that: 1) teacher attitudes have a clear and strong correlation to the use of LSA data; 2) teachers with positive attitudes about the data are most likely to make use of them; 3) attitudes variables only have significant impact on teaching (to) the curriculum strategies and not teaching to the test; and 4) provincial variations in testing policy go a long way to explain the variation in the results for 'teaching to the test' regression results.

While it has been somewhat established in the literature previously that attitudes do make a difference in regard to using LSA data (Brown 2004; Brown, Lake and Matters 2011), the quantitative aspects of this study bring that into better focus. It is also true that any Canadian study (this one being the largest in scale of its type known to the researcher) would see LSA policy in terms of testing that has limited professional consequence for teachers. The development of assessment policy that intends to change teaching practice appears to get better traction by addressing attitudes (and to a lesser extent supports) than with the use of incentives (see table 2).

This should be a word of guidance for both policy-makers and those central and school-based administrators tasked to oversee the implementation of LSA-based instructional improvements.

Another striking finding that comes from this analysis is that provincial variations in testing policy and test design are quite telling in terms of using teaching to the test strategies. Keeping in mind that stakes in the Canadian context apply more to students than to teachers, we see that in provinces that do not have high-stakes exit exams (Nova Scotia and Saskatchewan are in this category) teachers show much less interest in simple score-improvement strategies. The one province that has a unique assessment design that does not involve students writing paper and pencil test (Manitoba does have four grade 12 provincial exit exams, but all earlier [formative] tests are teacher-completed outcome checklists) also shows much less teaching to the test. We can see that the nature of the assessments administered in terms of both design and stakes matters when seen in light of both teacher attitudes and reactivity effects. When designing or re-tooling provincial LSA programs, education ministers would do well to note what effects their colleagues' programs have on teaching. There exists a national forum (the Council of Ministers of Education, Canada - CMEC) for just such an open discussion.

A third finding of some value is the fact that negative attitudes about testing do not correlate significantly with either type of reactivity. Teachers who see little value in large-scale assessment, for whatever principled or practical reason, are also not very likely to use the results in any way to adjust their instructional methods. It might be thought that teaching to the test is an obvious way to live up to accountability standards without necessarily having to take on the extra time and work entailed in teaching (to) the curriculum strategies. Negative attitudes about testing result most commonly in not using the data, which may be the most wasteful outcome from the perspective of the ministries who develop and fund their wide-spread use. It would be prudent to hold expensive and sometimes disruptive accountability testing regimes accountable for data that are not used or not used according to ministry guidelines.

There remains several papers to come that will pick up lines of inquiry not yet fully examined from this set of national data. Test design and results data, incentives and supports variables will all be given a thorough hearing. It is hoped that together this series of papers will shed some much needed light on the current and evolving practice of large-scale assessment in Canada and make clearer what policy options lead to the most desirable educational outcomes. This paper has shown that effective policy implementation depends in large part upon having willing and committed personnel in place in our front-line educational positions to carry through in the true spirit of a given assessment program with a high level of fidelity. Considering that teaching to the test is much more prevalent within Canadian classrooms than teaching (to) the curriculum, there remains much work to be done to draw teachers onside with large-scale assessment if it is to be useful to them, or used by them in constructive ways.

References

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1-74. http://dx.doi.org/10.14507/epaa.v10n18.2002

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

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. http://dx.doi.org/10.1080/09243450500114819

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

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27,* 210–220. http://dx.doi.org/10.1016/j.tate.2010.08.003

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297-312. http://dx.doi.org/10.1037/h0040950

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27. http://dx.doi.org/10.1111/j.1745-3992.2001.tb00072.x

Cobb, P., & Jackson, K. (2015). Supporting teachers' use of research-based instructional sequences. *ZDM Mathematics Education, 47, 1027-1038*. http://dx.doi.org/10.1007/s11858-015-0692-5

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

Copp, D. T. (2016). *Teaching to the test: A mixed methods study of instructional change from large-scale testing in Canadian schools*. Manuscript submitted for publication.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173–206. http://dx.doi.org/10.1080/15366367.2011.626729

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education 104*(2), 289-319. http://dx.doi.org/10.1111/j.1744-7984.2005.00034.x

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*(040302), 1-26.

Fehr, D. (2008). Financial industry certification preparation and "teaching to the test". *Journal of Economics and Finance Education, 7*(1), 1-6.

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

Flick, U. (2006). An introduction to qualitative research (3rd ed.). London: Sage Publications.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: a construct validation. *Journal of Educational Psychology, 76*(4), 669-682. http://dx.doi.org/10.1037/0022-0663.76.4.569

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized assessment test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7. http://dx.doi.org/10.3102/0013189X020005002

Hamilton, L. S., & Berends, M. (2006). Instructional practices related to standards and assessments. Santa Monica, CA: RAND. Hanushek, E. A., & Raymond, M. E. (2002, June.). Improving educational quality: How best to evaluate our schools? Lecture presented at Education in the 21st Century: Meeting the Challenges of a Changing World, Boston, USA.

Jacob, B. A. (2002, June). Test-based accountability and student achievement gains: Theory and evidence. Lecture presented at Taking Account of Accountability: Assessing Politics and Policy, 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. http://dx.doi.org/10.2307/2392366

Kitiashvili, A. (2014). Teachers' attitudes toward assessment of student learning and teacher assessment practices in general educational institutions: The case of Georgia. *Improving Schools, 17*(2), 163-175. http://dx.doi.org/10.1177/1365480214534543

Koretz, D., & Jennings, J. L. (2010). The misunderstanding and use of data from educational tests. Lecture presented at The Process of Data Use, The Spencer Foundation, Chicago, IL.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources, 37*(4), 752-777. http://dx.doi.org/10.2307/3069616

Koretz, D. (2009). Implications of current policy for educational measurement. Paper presented at the Center for K-12 Assessment & Performance Management, Princeton, NJ.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16. http://dx.doi.org/10.3102/0013189X029002004

Loughran, J. J. (2002). Effective reflective practice: In search of meaning in learning about teaching. *Journal of Teacher Education, 53*(33), 33-43. http://dx.doi.org/10.1177/0022487102053001004

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal, 91*(1), 79-87. http://dx.doi.org/10.2307/821659

Marion, S., & Leather, P. (2015). Assessment and accountability to support meaningful learning. *Education Policy Analysis Archives, 23*(9), 1-19. http://dx.doi.org/10.14507/epaa.v23.1984

McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

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

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

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.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education, 27*(2), 472-482. http://dx.doi.org/10.1016/j.tate.2010.09.017

Santos, J. R. (1999, April). Cronbach's Alpha: A Tool for Assessing the Reliability of Scales. Retrieved March 2, 2015, from http://www.joe.org/joe/1999april/tt3.php?ref

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

Schifter, C. C., Natarajan, U., Ketelhut, D. J., & Kirchgessner, A. (2014). Data-driven decision making: Facilitating teacher use of student data to inform classroom instruction. *Contemporary Issues in Technology and Teacher Education, 14*(4), 419-432.

Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation, 37*, 49-34. http://dx.doi.org/10.1016/j.stueduc.2011.03.008

Simner, M. L. (2000, Apr. 1). A joint position statement by the Canadian Psychological Association and the Canadian Association of School Psychologists on the Canadian Press coverage of the province-wide achievement test results. *Canadian Psychological Association*. Retrieved Aug. 9, 2012, from http://www.cpa.ca/documents/joint_position.html

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

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

Stock, J. H., & Watson, M. W. (2007). Introduction to econometrics (2nd ed.). Boston: Pearson/Addison Wesley.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. Retrieved March 2, 2015 from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4205511/ http://dx.doi.org/10.5116/ijme.4dfb.8dfd

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region. Melbourne: Australian Council for Educational Research.

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35*. http://www.umanitoba.ca/publications/cjeap/articles/volante.html

Volante, L. (2005). Accountability, student assessment, and the need for a comprehensive approach. *International Journal for Leadership in Learning, 9*. http://files.eric.ed.gov/fulltext/EJ985390.pdf

Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy, 58*, 1-21. http://umanitoba.ca/publications/cjeap/articles/volante_educational%20_quality.html

Volante, L., & Cherubini, L. (2007). Connecting educational leadership with multi-level assessment reform. *International Journal for Leadership in Learning, 11*(12). http://www.ucalgary.ca/~iejll/

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), 1-31. http://dx.doi.org/10.14507/epaa.v20n25.2012

Appendix

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

[1] Graduation requirement (must be passed)

[2] Graduation requirement only when teacher not accredited

[3] Mark on exam assigned a designated value of final grade

[4] Re-write of graduation requirement exam

[5] New in the 2013-2014 school year

[6] Student *must* write both EF and SS

[7] Suspended in the 2012-2013 and 2013-2014 school years

[8] Language Arts exams may include reading and/or writing components

EF - Core English or French

M - Mathematics

S - Science

SS - Social Studies

O - Other

Figure 4: Grades and subjects tested in Canadian provinces

Table 4

Ranking Canadian provinces based on average reactivity scores from survey data

	Prov.	Avg. score	SD	Prov.	Avg. score	SD	Prov.	Avg. score	SD	
Rank	Teaching (to) the curriculum			Teaching to the test			Net Reactivity			*n*
1st	AB	2.95	1.08	ON	-3.73	1.17	NS	0.21	1.40	46
2nd	NB	2.84	1.05	QC	-3.52	1.11	PEI	-0.07	1.14	39
3rd	PEI	2.82	1.16	NL	-3.53	0.80	SK	-0.21	1.17	37
4th	QC	2.78	1.08	AB	-3.45	1.15	MB	-0.3	1.20	55
5th	NL	2.72	0.90	NB	-3.19	1.03	NB	-0.35	1.38	29
6th	MB	2.41	1.11	BC	-3.09	1.21	AB	-0.5	1.36	48
7th	NS	2.34	1.17	PEI	-2.90	1.03	QC	-0.74	1.37	51
8th	ON	2.26	1.25	MB	-2.70	1.28	NL	-0.81	1.28	34
9th	SK	2.17	1.17	SK	-2.38	1.17	BC	-1.28	1.16	29
10th	BC	1.81	0.95	NS	-2.14	1.31	ON	-1.46	1.30	50
	CANADA	2.50	1.15	CANADA	-3.04	1.24	CANADA	-0.54	1.37	418

Note: All respondents' scores were totalled, and divided by *n* to provide an average score. Net reactivity is calculated by adding the two other measures.

Appendix (cont'd)

Survey Outline

This survey was written for Survey Monkey which allowed it to be emailed to teachers and results compiled electronically. The bracketed values beside the survey responses indicate binary 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.

Background

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 Data

14. Which tests does your class(es) write? 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)

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

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

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: 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

All responses in the following sections are scored -1 for 'Disagree', 1 for 'Agree', and 0 for 'Neither agree nor disagree.'

33. Provincial testing: (School Accountability)	a. Is a good way to evaluate a school b. Is an accurate indicator of a school's quality
34. Provincial testing: (Student Accountability)	a. Determines if students meet qualification requirements b. Makes parents better aware of student growth c. Selects students for education / employment opportunities
35. Provincial testing: (School Improvement)	a. Identifies student strengths and weaknesses 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

All responses in the following sections are scored 1 for 'Disagree', -1 for 'Agree', and 0 for 'Neither agree nor disagree.'

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

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])