Teacher-based reactivity to provincial large-scale assessment in Canada

Teacher-based reactivity to provincial large-scale assessment in Canada
By: Derek Copp

Summary, conclusions and recommendations

This brief paper summarizes the findings of the study, draws general conclusions, and finally makes recommendations for policy makers regarding provincial assessments. The layout of the paper is as follows:

- Section 1 examines the scale and purpose of the study by looking at the research question and also how and when data were collected
- Section 2 presents the reactivity results examined without the inclusion of independent variables
- Section 3 shows how test design/data independent variables affect teachers' use of large-scale assessment (LSA) data
- Section 4 submits results from attitudes-based independent variables and their influence on reactivity
- Section 5 looks at the effects of provided supports on reactivity effects
- Section 6 examines the influence of incentives on reactivity effects
- Section 7 presents reactivity results in light of background factors such as teacher experience, class size, and school setting
- and Section 8 includes recommendations based on the study's findings

1 Parameters of this study

Through surveys and interviews this study has examined how (or if) teachers change their instructional practices based on the results data from large-scale testing done in all Canadian provinces. While the subjects and the grade levels tested vary, some general conclusions about the effects of testing on instruction have been drawn. All education ministries indicate in their assessment policy literature that these tests are intended to improve teaching and/or to help identify students who need instructional interventions of some sort. To ask teachers how their teaching has changed, how 'reactive' they are to provincial assessment data, is a program evaluation question considered in light of the ubiquitous policy choice to test large numbers of students annually in Canadian public schools. According to the ministries' own literature, reactivity is an expectation for professional staff (see Chapter 1).

Reactivity is a framework for examining how people react in situations in which they are being externally evaluated. Large-scale assessments are used across Canada as a form of external evaluation of both teachers and schools. The changing of teaching practices is expected and normal. *How* they change depends on several factors. Positive reactivity is defined in this dissertation to include those instructional practices and strategies that are both ethical and broaden the number and variety of outcomes presented to students. Negative reactivity is defined in this dissertation as a set of instructional strategies and practices which are either unethical or reduce the number or variety of outcomes presented to students. Both of these definitions rest on the foundations of the Saskatchewan Teachers' Federation Code of Professional Conduct (found in Annex 1).

Data were collected in a nationwide survey of teachers in Canada to gauge the amount and type of reactivity effects which result from provincial testing as well as responses regarding the independent variables which may go some way to explain them. Using these data, two other metrics of reactivity were created and examined: (a) total reactivity which adds all reactivity effects to show how much of teachers'

practices are affected by assessments; and (b) net reactivity which determines the overall tendency of the data after subtraction of negative effects from the positive.

There were several lines of inquiry that were followed in the survey to determine their influence on reactivity effects: (a) teachers were asked their opinions of the test design, and of the results data; (b) they were asked their attitudes about large-scale testing in general; (c) inquiries were made about the supports provided to help them use the data; (d) they were asked what incentives were in place to encourage data use; and (e) general background questions were asked to gauge the representativeness of the sample and to determine if any of these factors were correlated with data use. Each of these lines of inquiry was discussed in a chapter of this study, after which the survey results were used to do OLS statistical regressions intended to uncover correlations between these independent variables and the dependent variable, teacher reactivity.

Follow up interviews were also conducted with different stakeholders in the educational hierarchy. It was important to hear more from teachers directly, but also to discuss assessment data use with principals and division-level staff. These are often the people who establish a culture of data use or who set the expectation (implicit or explicit) that the data get used to inform instructional practices. These perspectives were used to complement the survey analyses according to variable-dependent coding.

2 Reactivity conclusions

Reactivity was measured by asking 10 questions about the instructional practices of teachers in response to the LSA results. These responses were on an ordinal scale (these being 'never,' 'sometimes', and 'always') which were then converted into a ratio scale (0, 0.5, and 1) and finally collated into national and provincial averages.

There is a wide variance in reactivity effects across Canadian provinces. While not all of this variance is attributable to the independent variables examined in this study, there are some general conclusions that can be drawn before these independent variables are themselves examined.

- **Teachers are in general strongly reactive to provincial assessment data**
- **Negative reactivity effects are dominant in 9 of 10 Canadian provinces**
- **Not all provinces have comparable types or degrees of reactivity effects**

These conclusions give policy makers some reason to feel pleased but also suggest that there is plenty of room for improvements to be made. The fact that teachers are generally reactive is good news – it is an expectation across Canada that the results from large-scale provincial assessments are seriously considered and that they point the direction to instructional and programming improvement.

Improvements are possible in that the degree to which the teaching population is reactive varies widely between provinces, and also in that the type of reactive effect employed is most commonly (by the terms defined in this dissertation) negative reactivity. Alberta teachers are the most reactive in Canada (scoring 6.39 on a scale that has a maximum of 10), yet they also are inclined towards negative reactivity effects (the net overall score is -0.50). Nova Scotia is the one and only province with net positive reactivity effects (the net score is 0.21), yet it is also the least reactive province (scoring 4.48 on the scale to 10 for total reactivity).

These preliminary results showed reactivity scores for survey respondents, but were not used here for the regression analysis: there were no measures of strength employed beyond these numbers themselves. The sections that follow include statistical analyses for different lines of inquiry and the several independent variables in each.

3 Test design and results conclusions

Test design was not a constant in this study – each province designs and distributes their own assessments. Nor are the data returned to teachers uniform in nature. Even so, the relative effect of design and data return variables on reactivity is the focus of this section.

- **None of the test design or data variables had a significant correlation with positive reactivity**
- **More aggregated and disaggregated data returned to teachers correlated with more negative effects**
- **For total reactivity effects, more aggregated and disaggregated data returned to teachers was significantly correlated, indicating that detailed results data makes likely more positive and negative reactivity effects**
- **Quite few and relatively weak provincial variations were in evidence**

These results seem to suggest that some factors cited by teachers as important factors in their decision-making process regarding LSA data did not correlate to using the data. There was no significant effect on reactivity results regardless of their opinions about test design, whether the data were clear, nor whether they felt prepared to act on the data. The types of data returned did make a difference in terms of both negative and total reactivity effects. Getting aggregated and disaggregated results seemed to increase negative reactivity effects more so than positive, but the just less-than-significant positive reactivity correlation was a contributing factor in the significant total reactivity result. Remarks in the literature related to the low assessment literacy of teachers coming out of pre-service programs may shed some light on this dynamic. [1]

Including provincial dummies, these correlations are quite strong, with R^2 values going from 14% (positive), to 18% (negative) and finally 16% (total).

4 Test attitudes conclusions

Those measures of teachers' attitudes about testing that were used in the survey were more strongly correlated with reactivity effects than opinions about the test design or the data returned. This seems to indicate that attitudes about testing are more important than opinions or critiques of the instruments used. It is also the case that there were wide variances between provinces on these measures.

- **3 variables (the use of data for student accountability, for school improvement, and agreement that there are more appropriate uses for the data) had significant and strong correlations with positive reactivity.[2]**
- **No variables had a significant effect on negative reactivity, while provincial variations were in evidence**
- **Despite the low level of correlation between attitudes and negative reactivity, total reactivity showed significant relationships for 2 variables**

So while strong beliefs about both school and student accountability were cited by teachers and higher level officials alike, neither of these factors had a large impact on reactivity effects. The more

[1] See, for example: Lukin, Bandalos, Eckhout & Mickelson, 2004; Hargreaves, Crocker, Davis, McEwen, Sahlberg, Shirley, & Sumara, 2009; Earl & Fullan, 2003.

[2] Using data for student accountability, for school improvement, and agreement that there are more appropriate uses for the data.

telling factors were the belief that LSAs can lead to school improvement, positive attitudes about testing in general, and the belief that the data could be put to appropriate use.

With the provincial dummies included, the variables in this section account for 23% of the variation in positive reactivity, 20% of the variation for negative reactivity, and 22% for total reactivity scores. It is clearly important in terms of reactivity for teachers to feel that LSAs can improve teaching and that they see the utility in these assessments.

5 Supports conclusions

Supports for the use of data are provided at all three jurisdictional levels: schools, divisions, and ministries. The number of available supports was asked, as was how helpful these supports were considered by teachers. Both the number and perceived quality differed widely across provincial jurisdictions.

- **The sharing of data and division-level supports are strongly and significantly correlated to positive reactivity effects**
- **No variables were significantly correlated with the use of negative reactivity practices while some provincial variation is evident**
- **Total reactivity mirrored the previous results and show strong significant correlations between sharing data and divisional-level supports with some provincial variances**

These results show that the sheer number of supports matters much less than a culture of sharing data and also the jurisdictional level from which supports come. Divisional supports were less common and less highly regarded by respondents, but only supports from the division level were tied to reactivity effects. There is a school-level factor that comes into play here, and that is sharing data. Schools with cultures that support and encourage data-sharing are more inclined towards positive reactivity effects.

The correlations from this chapter are strong, explaining fully 22% of positive reactivity scores, 19% of negative reactivity scores, and 21% of the variance in total reactivity scores.

6 Incentives conclusions

Incentives are generally built into assessment policies, even if they are as simple as the explicit expectation that data will be used and follow up on that expectation. The perceived amount of pressure and level of stakes for teachers were also part of this line of inquiry.

- **Perceived pressure had the strongest correlation to positive reactivity, while follow up and perceived stakes had lesser significance**
- **Perceived pressure had a strong and significant impact on negative reactivity**
- **Perceived pressure was, as above, the strongest factor in accounting for total reactivity effects while results awareness had a less telling effect**

It is clear from the results that the amount of pressure teachers feel is a key determinant in their use of LSA data which may bolster the opinion that 'incentives work.' Yet they work to create both positive and negative effects. Perceived personal or professional stakes seemed less important and were significant only to positive effects. Nor were the expectations to use data or the follow up on that expectation as significant as one might imagine (since these might been seen as the more obvious aspects of the pressure teachers feel is being applied). The results awareness result was significant to only

negative and total reactivity correlations, but clearly being aware of the results has a significant impact on the use of the data in either a pro-active or re-active sense.

The variables examined had a strong influence on the variances in positive reactivity scores (24%), negative reactivity scores (18%), and also total reactivity scores (23%).

7 Background variables conclusions

Background variables were not expected to show strong or significant correlations with reactivity effects, but since the data were already gathered to look at the representativeness of the survey sample, it made sense to do statistical analysis with them if only to pre-empt questions around these considerations.

- **Class size and age both have small but significant correlations with positive reactivity**
- **The grade level taught has strong and significant correlations to negative reactivity**
- **Age, grade taught and class size have weaker effects on total reactivity while some provincial variations are evident**

Only one of the correlations above was significant at the $p < 0.01$ confidence level. The results from the provinces seem to point to the fact that high stakes exit examinations (all of these given at grade 11 or 12 level, and thus an important factor in the 'grade taught' data) were the difference between this correlation being significant or not.

The relationship between teacher age and less reactivity was not expected, but was not particularly strong. It is perhaps unkind to write this off as being a result of 'trying to teach an old dog new tricks' but the fact that the experience variable (which should theoretically be covariant with the age variable) is not covariant makes the result more confounding. It is likely that high data usage by younger teachers would account in part for this result.

The effects of these variables are relatively strong, explaining only 12% of positive reactivity variances, but 21% of negative reactivity effects, and 17% of the total reactivity variances. Provincial variances are a major component of these high R^2 values**.** It is interesting to note that none of these factors is nearly as predictive in terms of positive reactivity as they are in terms of negative effects.

8 Recommendations

Results from surveys and triangulated interview data have made clear some aspects of Canadian large-scale assessment policies that otherwise might have remained unclear. Suppositions and statistical correlations are different things, and only the latter is a solid foundation for making or amending educational policy. With this in mind these conclusions do have practical policy applications that may serve to strengthen or at least clarify the purposes of provincial assessment programs.

1. Start with education about testing

The most important single factor in this study seems to be the attitude of the teachers about testing. Some interview subjects were very cynical and dismissive of the assessments, but others saw the value in even a flawed metric if the data were openly shared and used with discretion and after consultation with other teachers and/or consultants.

Seeing that attitudes about testing are this important, provincial education ministries and school divisions cannot leave it to random chance, pre-service training (as above in section 8.3, from Lukin et al., 2004; Hargreaves et al., 2009; Earl & Fullan, 2003, and also: Stiggins, 2002; Volante, 2004) or somewhat intangible personal variables (like charismatic leadership, for example) to determine whether their testing

policies are effective or not. Divisions seem to have a key role in effective professional development, and the ministries must support this. Professional growth through effective provincial assessment is only possible where school leaders and early-adopters have been converted to the belief that data use can be helpful, that specific tests can have positive impacts, and that there are benefits to teachers and students at the classroom level. Abstractions about 'benefits to policy-makers' are not sufficient to sway current instructional practices.

2. Make clear the differences between positive and negative uses of the data

One of the most troubling aspects of this study was the apparent unwillingness or inability of teachers to discern between positive and negative reactivity effects. While it is the author's own reactivity model that defines this distinction, there is some agreement in the research community about those instructional practices which provide a broader range of outcomes for students and those which diminish them. The model used here takes existing, well- documented beliefs about professional teaching practices and puts them into survey questions which permit the quantification of these data.

Total reactivity scores were quite high across the nation, showing that teachers *do* change their instruction based on LSA data, yet it seemed to not to make much difference whether these changes were ethical (or not) and provided a wide range of educational outcomes for students (or not). Respondents to the survey were inclined overwhelmingly to negative reactivity practices (while they were not identified in the survey as such). This is concerning. It should follow directly from recommendation #1 that if you want instructional change you must show the benefits of such changes and also make clear what kind of changes can and should be made. Based on the survey data results, the differences between test-based, narrow, assessment-focused practices and those that are learning-based and broaden curriculum outcomes must be made much more explicit to teachers in policy documents, in provided professional development, and whenever LSAs are discussed and used by the education sector.

3. Provide appropriate level support

In order to implement such large-scale policy goals, provinces need to be aware of which methods of implementation provide the most effective returns. It is clear that teachers feel that they need supports in many cases (especially for teachers unfamiliar with the tests and data, and when testing policies or tests change). The results from this study show that division-level support, while neither as common nor as well-received as school-level support, is correlated with improved reactivity outcomes and specifically positive reactivity effects. In this way, recommendation #3 follows from the implementation of recommendations #1 and #2. Teachers will first need some convincing (#1), and then some detailed information about positive instructional change (#2), but this in-servicing should ideally come from the division level (#3).

With the guidance and the support of the ministry, divisions must provide high level, universally-attended, and ongoing professional development for teachers who give LSAs in their classrooms. To ensure policy fidelity, the same message must go out to all teachers and administrators to create a culture of sharing and using data in educationally defensible ways.

4. High stakes exit examinations are prime candidates for negative effects

Of all the different assessment policy variations noted in this study, only one appears to have a strong link to negative reactivity but not to positive reactivity, and that is high stakes exit exams. Any discussion of high stakes exit exams is also a discussion of the increased pressure that these exams generate. High school exit exams seem almost purpose-built to meet the conditions of negative reactivity

with the amount of pressure applied to teachers (from students, parents, and administrators), the importance of these results for future studies and graduation (engaging 'painted into a corner' behaviours from teachers), and the secrecy shrouding the tests to try to prevent inappropriate preparation or instruction.

All of the fail-safe measures do not prevent classroom instruction from being primarily focused on the high stakes test. It is unlikely that instruction in any subject that has a test of this nature at its conclusion can avoid: (a) the assessment parameters becoming the course parameters; (b) the means of assessing used on the LSA becoming the classroom's chosen method of assessing; (c) teachers focusing on what they believe will be on the exam; and (d) students striving to learn only those things that are on the final test. These are all negative reactivity effects. Campbell (1957) noted that people react to being observed or evaluated, and these reactions, whether consciously or unconsciously, have an impact on the objectivity of that measurement. Teachers react to LSAs because they are being indirectly evaluated by their students' results. This is inevitable. Yet high stakes tests tend to promote negative reactivity, even where (or perhaps because) secrecy surrounds the test.

5. An assessment needs to have a clear and manageable purpose(s)

This study can be seen as a program evaluation project, comparing the stated policy goals for provincial assessments in Canada to their practical effects in classrooms. The author's perspective is that of a teacher, so this seemed an important comparison to make as it relates to the day-to-day work expectations for many professional educators.

It seems clear from numerous interview respondents, though, that very few people at any level in the educational hierarchy take very seriously ***all*** of the numerous, diverse, and sometimes conflicting purposes that provincial ministries have set for their own assessment systems. It seems that policy makers have bitten off more than it is either practical or wise to chew. An assessment that is suited to individualized data on student weaknesses is not necessarily very good at providing data to compare classrooms or schools; and an assessment that provides data to monitor adherence to the curriculum does not always have the information a classroom teacher would need to make instructional improvements. Add in public accountability functions, graduation requirements, curriculum adherence, and improving central and local data-based decision making and one can see that a single annual assessment might not be sufficient to address the multiple purposes with which it has been tasked.

Education ministries and departments should consider this well. An assessment that is expected to do so many things might do none of them particularly well. Going down the 'multi-purpose assessment' path often means sacrificing the quality of the data for one or more of these purposes. Therefore the mandate for provincial assessment should be narrowly focused on the data needs of provincial level bureaucracy, while other purpose-built tests (from the division-level or even the school-level) could be employed for other data needs.

These five general recommendations are unlikely to be controversial. They do, on the other hand, have the support of a large data set collected by the researcher in the only nation-wide study of reactivity effects in Canadian large-scale provincial testing. The statistical analyses show some strong and important correlations between assessment policy parameters and how the data are practically employed. So, in the end, these uncontroversial recommendations might serve as a roadmap to avoid assessment policies that could hamper instructional improvements, cost countless taxpayer dollars, and alienate teaching professionals.