

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

© 2015 Derek Copp

Cover picture by Jordan Junek

ISBN: 978 90 8666 371 2

Publisher: Boekenplan, Maastricht
www.boekenplan.nl

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form, or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission in writing, from the author.

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

DISSERTATION

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus Prof. dr. L.L.G. Soete,
in accordance with the decision of the Board of Deans,
to be defended in public on Wednesday 17 June 2015, at 14:00 hrs.

by

Derek Copp

Supervisor

Prof. dr. Jo Ritzen

Prof. dr. Louis Volante (Brock University)

Prof. dr. Jan van den Brakel

Assessment Committee

Prof. dr. Lex Borghans (Chair)

Prof. dr. Jaap Dronkers

Prof. dr. Henriette Maassen van den Brink

Prof. dr. Sally Thomas, University of Bristol, Graduate School of Education

Prof. dr. Michael T. Nettles, Educational Testing Service, Policy Evaluation and Research Center

Contents

Acknowledgements	pg. 1
<u>Chapter 1</u>	
1.1 Introduction	pg. 3
1.1.1 The roots of accountability	pg. 3
1.1.2 International pressures	pg. 4
1.1.3 Other purposes of assessments	pg. 5
1.1.4 The limits of tests	pg. 6
1.1.5 Changes in instruction	pg. 7
1.1.6 Opportunity cost	pg. 7
1.2 Research questions	pg. 8
1.3 Motivation	pg. 9
1.4 Canadian context	pg. 11
1.4.1 National demographic information	pg. 11
1.4.2 Provincial demographic information	pg. 15
1.5 Charts	pg. 31
<u>Chapter 2</u>	
2.1 Introduction	pg. 32
2.1 Mixed methods	pg. 32
2.2 Surveys	pg. 35
2.3 Survey sampling	pg. 37
2.4 Other potential errors	pg. 39
2.5 Survey data analysis	pg. 41
2.6 Interviews	pg. 45
2.7 Canadian context	pg. 50
2.8 Conceptual framework	pg. 51
2.9 Charts	pg. 53
<u>Chapter 3</u>	
3.1 Introduction	pg. 54
3.2 Theoretical framework	pg. 54
3.2.1 Other reactivity studies	pg. 55
3.2.2 Educational reactivity	pg. 55
3.3 Reactivity model	pg. 61
3.4 Reactivity survey results	pg. 64
3.4.1 National data	pg. 65
3.4.2 Provincial results	pg. 68
3.5 Conclusions	pg. 89
3.6 Charts and tables	pg. 92

Chapter 4

4.1	Introduction	pg. 98
4.2	Literature review	pg. 98
4.2.1	Policy-level considerations	pg. 98
4.2.2	Classroom-level considerations	pg. 102
4.3	Preliminary hypotheses	pg. 112
4.4	Results from surveys – tests and data	pg. 112
4.4.1	National results	pg. 113
4.4.2	Provincial results	pg. 115
4.5	Correlation analysis – tests and data	pg. 138
4.6	OLS regressions – tests and data results	pg. 139
4.6.1	Regression analysis	pg. 139
4.6.2	Residual analysis	pg. 145
4.7	Results from surveys – test attitudes	pg. 145
4.7.1	School accountability results	pg. 147
4.7.2	Student accountability results	pg. 152
4.7.3	School improvement results	pg. 154
4.7.4	Negative test attitudes results	pg. 158
4.7.5	Appropriate uses results	pg. 160
4.8	Correlation analysis – test attitudes	pg. 163
4.9	OLS regressions – test attitudes	pg. 164
4.9.1	Regression analysis	pg. 164
4.9.2	Residual analysis	pg. 170
4.10	Conclusions	pg. 172
4.11	Charts and tables	pg. 174

Chapter 5

5.1	Introduction	pg. 192
5.2	Literature review	pg. 192
5.3	Preliminary hypothesis	pg. 200
5.4	Results from surveys - supports	pg. 201
5.4.1	National results	pg. 201
5.4.2	Provincial results	pg. 203
5.5	Correlation analysis - supports	pg. 225
5.6	OLS regressions – supports	pg. 226
5.6.1	Regression analysis	pg. 226
5.6.2	Residual analysis	pg. 230
5.7	Conclusions	pg. 232
5.8	Charts and tables	pg. 235

Chapter 6

6.1	Introduction	pg. 242
-----	--------------	---------

6.2 Literature review	pg. 242
6.3 Preliminary hypothesis	pg. 248
6.4 Results from surveys – incentives	pg. 248
6.4.1 National results	pg. 249
6.4.2 Provincial results	pg. 252
6.5 Correlation analysis – incentives	pg. 274
6.6 OLS regressions - incentives	pg. 275
6.6.1 Regression analysis	pg. 275
6.6.2 Residual analysis	pg. 281
6.7 Conclusions	pg. 283
6.8 Charts and tables	pg. 285

Chapter 7

7.1 Introduction	pg. 291
7.2 Literature Review	pg. 291
7.3 Preliminary hypothesis	pg. 300
7.4 Results from surveys	pg. 300
7.5 Correlation analysis – background factors	pg. 309
7.6 OLS regressions – background variables	pg. 310
7.6.1 Regression analysis	pg. 310
7.6.2 Residual analysis	pg. 316
7.7 An inquiry into subject area qualifications	pg. 316
7.8 Conclusions	pg. 320
7.9 Final words	pg. 321
7.10 Charts	pg. 324

Chapter 8

8.1 Parameters of this study	pg. 326
8.2 Reactivity conclusions	pg. 327
8.3 Test design and results conclusions	pg. 328
8.4 Test attitudes conclusions	pg. 329
8.5 Supports conclusions	pg. 330
8.6 Incentives conclusions	pg. 330
8.7 Background variables conclusions	pg. 331
8.8 Recommendations	pg. 332

Bibliography

	pg. 336
--	---------

	pg. 383
--	---------

Appendix 1: STF Code of Professional Competence	pg. 383
-------------------------------------------------	---------

Appendix 2: Teachers survey	pg. 384
-----------------------------	---------

Appendix 3: Interview guide	pg. 394
-----------------------------	---------

Appendix 4: Interview coding key	pg. 395
----------------------------------	---------

MGSOG Dissertation Series Titles	pg. 396
----------------------------------	---------

Chapter 1

Figure 1.1: Provincial testing in Canada	pg. 13
Figure 1.2: Purposes of provincial testing in Canada	pg. 14
Figure 1.3: First Nations populations in Canada	pg. 18
Figure 1.4: Francophone populations in Canada	pg. 20
Figure 1.5: First Nations populations as a % of province	pg. 24
Figure 1.6: Francophone populations as a % of province	pg. 27
Figure 1.7: Student-educator ratios for Canadian provinces	pg. 31

Chapter 2

Table 2.1: Number of responses to nation-wide teacher survey	pg. 41
Figure 2.2: Summary of the methodology literature	pg. 46
Figure 2.3: Respondents to surveys and interviews	pg. 50
Figure 2.4: Conceptual framework	pg. 52
Figure 2.5: Sampling method	pg. 53

Chapter 3

Figure 3.1: Summary of reactivity literature	pg. 57
Figure 3.2: Positive and negative reactivity survey questions	pg. 62
Table 3.3: Ranking Canadian provinces based on reactivity effects	pg. 67
Figure 3.4: Positive reactivity scores	pg. 92
Table 3.5: Distribution analysis for national data from figure 3.4	pg. 92
Figure 3.6: Aggregated data from figure 3.4 (three groupings)	pg. 93
Figure 3.7: Negative reactivity scores	pg. 93
Table 3.8: Distribution analysis for national data from figure 3.7	pg. 94
Figure 3.9: Aggregated data from figure 3.7 (three groupings)	pg. 94
Figure 3.10: Total reactivity scores	pg. 95
Table 3.11: Distribution analysis for national data in figure 3.10	pg. 95
Figure 3.12: Aggregated data from figure 3.10 (three groupings)	pg. 96
Figure 3.13: Net reactivity scores	pg. 96
Figure 3.14: Detailed national data for net reactivity	pg. 97
Table 3.15: Distribution analysis for national data from fig. 3.14	pg. 97

Chapter 4

Figure 4.1: Ethical and unethical educational practices	pg. 103
Figure 4.2: Summary of test design, data and attitudes literature	pg. 105
Table 4.3: Correlation matrix for test design and data variables	pg. 138

Table 4.4: Positive reactivity, test design and results variables	pg. 140
Table 4.5: Negative reactivity, test design and results variables	pg. 142
Table 4.6: Total reactivity, test design and results variables	pg. 144
Figure 4.7 Residual analysis for tests and data positive reactivity	pg. 146
Table 4.8: Correlation matrix for test attitude variables	pg. 163
Table 4.9: Positive reactivity, test attitude variables	pg. 165
Table 4.10: Negative reactivity, test attitude variables	pg. 167
Table 4.11: Total reactivity, test attitude variables	pg. 169
Figure 4.12: Residual analysis for test attitudes total reactivity	pg. 171
Figure 4.13: Percentage of teacher getting same-year results	pg. 174
Figure 4.14: Percentage of teachers getting agg./disagg. data	pg. 174
Figure 4.15: Percentage of teachers consider items appropriate	pg. 175
Figure 4.16: Teacher opinions of selected-response items	pg. 175
Figure 4.17: Teacher opinions of short constructed-response items	pg. 176
Figure 4.18: Teacher opinions of long constructed-response items	pg. 176
Figure 4.19: National teachers' opinions of LSA item types	pg. 177
Figure 4.20: How LSA results are shared	pg. 177
Figure 4.21: National/provincial data on understanding of results	pg. 178
Figure 4.22: Teachers ability to act on results data	pg. 178
Figure 4.23: School accountability responses	pg. 179
Figure 4.24: Student accountability responses	pg. 179
Figure 4.25: School improvement responses (test-giving teachers)	pg. 180
Table 4.26: Distribution analysis for figure 4.25 data	pg. 180
Figure 4.27: School improvement (non-test-giving teachers)	pg. 181
Table 4.28: Distribution analysis for figure 4.28 data	pg. 181
Figure 4.29: School improvement responses (all teachers)	pg. 182
Table 4.30: Distribution analysis for figure 4.27 data	pg. 182
Figure 4.31: Test attitudes responses (test-giving teachers)	pg. 183
Table 4.32: Distribution analysis for figure 4.31 data	pg. 183
Figure 4.33: Test attitudes responses (non-test-giving teachers)	pg. 184
Table 4.34: Distribution analysis for figure 4.33 data	pg. 184
Figure 4.35: Test attitudes responses (all teachers)	pg. 185
Table 4.36: Distribution analysis for figure 4.35 data	pg. 185
Figure 4.37: Appropriate uses responses (test-giving teachers)	pg. 186
Table 4.38: Distribution analysis for figure 4.37 data	pg. 186
Figure 4.39: Appropriate uses responses (non-test-giving teachers)	pg. 187
Table 4.40: Distribution analysis for figure 4.39 data	pg. 187
Figure 4.41: Residual analysis for tests and data negative reactivity	pg. 188
Figure 4.42: Residual analysis for tests and data total reactivity	pg. 189
Figure 4.43 Residual analysis for test attitudes positive reactivity	pg. 190
Figure 4.44 Residual analysis for test attitudes negative reactivity	pg. 191

Chapter 5

Figure 5.1: Summary of supports literature	pg. 195
Table 5.2: Correlation matrix for supports variables	pg. 226
Table 5.3: Positive reactivity, supports variables	pg. 227
Table 5.4: Negative reactivity, supports variables	pg. 229
Table 5.5: Total reactivity, supports variables	pg. 231
Figure 5.6: Residual analysis for positive reactivity regressions	pg. 233
Figure 5.7: Sharing of data (test-giving teachers)	pg. 235
Figure 5.8: Sharing of data comparison (all teachers)	pg. 235
Figure 5.9: Supports and the jurisdictions that provide them	pg. 236
Figure 5.10: Most commonly provided supports	pg. 236
Figure 5.11: Jurisdictional provision of supports	pg. 237
Figure 5.12: Helpfulness of school supports	pg. 237
Figure 5.13: Helpfulness of divisional supports	pg. 238
Figure 5.14: Helpfulness of ministry supports	pg. 238
Figure 5.15: Helpfulness of supports – aggregate data	pg. 239
Figure 5.16: Residual analysis for negative reactivity regressions	pg. 240
Figure 5.17: Residual analysis for total reactivity regressions	pg. 241

Chapter 6

Figure 6.1: Summary of incentives literature	pg. 244
Table 6.2: Correlation matrix for incentives variables	pg. 275
Table 6.3: Positive reactivity, incentives variables	pg. 276
Table 6.4: Negative reactivity, incentives variables	pg. 278
Table 6.5: Total reactivity, incentives variables	pg. 280
Figure 6.6: Residual analysis for total reactivity regressions	pg. 282
Figure 6.7: Perceived expectations to use LSA data	pg. 285
Figure 6.8: Perceived follow up on expected use of data	pg. 285
Figure 6.9: Class-level results awareness	pg. 286
Figure 6.10: School-level results awareness	pg. 286
Figure 6.11: Division-level results awareness	pg. 287
Figure 6.12: Overall results awareness	pg. 287
Figure 6.13: Pressure reported by test-giving teachers	pg. 288
Figure 6.14: Perceptions of pressure for non- and test-giving	pg. 288
Figure 6.15: Perceived level of stakes	pg. 289
Figure 6.16: Perceptions of stakes for non- and test-giving teachers	pg. 289
Figure 6.17: Perceived stakes for non- and test-giving teachers	pg. 290
Figure 6.18: Residual analysis for positive reactivity regressions	pg. 291

Figure 6.19: Residual analysis for negative reactivity regressions pg. 292

Chapter 7

Figure 7.1: Summary of background factors literature	pg. 296
Figure 7.2: National and sample age data comparison	pg. 301
Figure 7.3: National and sample sex data comparison	pg. 302
Figure 7.4: School setting as reported by respondents	pg. 304
Figure 7.5: Grade levels taught as reported respondents	pg. 305
Figure 7.6: Years of teaching experience	pg. 306
Figure 7.7: School size as reported by survey respondents	pg. 307
Figure 7.8: Average class sizes as reported by survey respondents	pg. 307
Table 7.9: Distribution analysis for figure 7.8	pg. 308
Figure 7.10: Qualifications data as reported by survey respondents	pg. 308
Table 7.11: Correlation matrix for background factors	pg. 310
Table 7.12: Positive reactivity, background factor variables	pg. 311
Table 7.13: Negative reactivity, background factor variables	pg. 313
Table 7.14: Total reactivity, background factor variables	pg. 315
Figure 7.15: Residual analysis for positive reactivity regressions	pg. 317
Figure 7.16: Teachers who give English LSAs	pg. 318
Figure 7.17: Teachers who give Mathematics LSAs	pg. 318
Figure 7.18: Teachers who give Science LSAs	pg. 319
Figure 7.19: Teachers who give Social Studies LSAs	pg. 320
Figure 7.20: Residual analysis for negative reactivity regressions	pg. 324
Figure 7.21: Residual analysis for total reactivity regressions	pg. 325

Acknowledgements

A recent graduate of the Maastricht GPAC² program visited my cohort in the first year of study and said something that stuck with me. Dr. Joe Abah said that doing a PhD was at its foundation quite a selfish pursuit. Despite any lofty personal future goals (such as research, policy planning or an academic career), the sheer volume of reading, writing and thought required in this line of study meant that many other things, important things, had to be sidelined. He was absolutely right about this. I have been lucky enough to have support from many quarters during these three intense years without which I may not have been able to complete this work. I doubt that either seeing the dissertation complete or any thanks I give here are reward enough for these freely given gifts of selfless generosity.

My family, and in particular my wife Meaghan, must top the list. When I started this PhD we had just one small child, and now that I am complete we have a much bigger boy plus two other children. Duncan, Katrijn and Sebona were in good hands while I was away for a couple of weeks a year and also several nights every week. Meaghan gets all the credit for raising our wonderful family in my absence. Both my mother and Meaghan's were also willing to visit and help with the childrearing in my absence, and this was a big help. My in-laws have been particularly supportive, and I would be remiss not to thank Terry and Jean, as well as my mother Joan.

At my day job, there was a professional team that, I'm sure, barely even noticed that I was away such is their willingness to take change in stride. My principal Gord Erhardt had more administrative duties on his plate with my absences. Terri Parsons deserves a special mention for taking on the task of proofreading what was then a 450 page draft for nothing but a thank you and a bottle of Baileys. They don't make bottles big enough to show my appreciation. Other staff at EHS were happy to assist with duties, coaching, and whatever else came up while I was away. And I will not forget the financial and moral support provided from the Good Spirit School Division. Time away from classes is taxing on the school system, but Dwayne Reeve, our director, never questioned the value of my pursuit.

The GPAC² team in Maastricht made every visit to the Netherlands a joy. There were also compelling academic discussions with first-rate specialists from numerous fields of endeavour – it was impossible to be bored at GPAC sessions. Mindel and Eddy welcomed us to the rigour of PhD level study. They pulled no punches and lessons learned from them stuck. Carlos, Güney, Saba and Charlotte covered the logistics of the program. Their good nature and humour made it feel like visiting old friends each time I came.

I also had the good fortune to have a supportive, encouraging, and uncompromising supervisory team. Dr. Jo Ritzen was my thesis promoter and brought both a wealth of knowledge and practical suggestions to my aid. I could not have asked for a better promoter. Dr. Louis Volante took a place on the team after the work had taken on much of its current shape, yet his advice on how to improve the work for academic publication and his insights into the topic itself were enlightening and welcome. Dr. Jan van den Brakel also starting working with me later in the project and I think I relied on his econometric expertise most of all. While I know my way around schools and have a fairly good handle on this topic, the analysis of the data was something I had to learn to do, and Jan was the unfortunate soul tasked with answering all of my questions. Thank you, Jan, for sharing some of your knowledge but especially for your patience.

I should mention that the other GPAC fellows were a limitless source of strength, experience and friendship. Ana, Corrine, Rafa, Camillo, Shellie, Casty, Hoda and Corinne all started with me in March 2012, and all of us have stuck through the hard moments to see our work progress. Without their kindness and enthusiasm, this would not have nearly so engaging a journey. Even those whose paths eventually diverged from GPAC, Rron, Alexis and Daniel, made the experience richer and I am thankful for the time they were there. I will really miss these twice-yearly trips to be with friends, to get lots of great ideas, and to try some of the local beer.

This whole study rests on the voluntary participation of teachers, and for this reason the professional educators across Canada deserve some recognition. Teachers only ever had a chance to participate if the school administrator agreed to distribute my survey, so principals and vice principals also earn a nod. Even those who said 'no' often did so for reasons I respect, so hat's off to them all. The staffs at several school districts and divisions were responsible for deciding whether or not to grant me access to school administrators, and I am indebted to them for allowing this study to be done at all.

I will admit now that I really had no idea what I was signing on for when I joined this program. A PhD seemed like simply the next logical step after a Master's, but it turned out to be much more. No preparations beyond a good academic standing and having good work habits are required for university study even up to the graduate level. Yet these criteria proved to be necessary but not sufficient for PhD work. I think that the combination of rigour and long-term nature of the PhD demands more than the more basic skill set in order to see success. You need, in the end, more support and forgiveness from your friends, your family and from those kinds spirits you meet along the road. To all of you, I give my sincere thanks.

Introduction to the topic

1.1 Introduction

Classroom teachers, at all grade levels and locations, see a significant amount of instructional time each year spent on non-instructional initiatives. Some of this 'lost time' is inevitable and can make the school experience more engaging for students. Standardized tests, mandated for use by all 10 Canadian provinces, are an ever-increasing part of this non-instructional time. From the standpoint of a teacher, the researcher has wondered if this lost time can be justified on the basis that it provides educational benefits. Since one of the main purposes of standardized tests is the improvement of instruction at the classroom level (Klinger, DeLuca & Miller, 2008), it warrants examination whether or not this is the case. If large-scale assessment results *are* used, are these data used well or used badly? The question that follows naturally in the Canadian context is which provincial models of testing can be shown to be the most successful at attaining positive instructional change?

This chapter is laid out in the following way: (a) the introductory section will discuss the pressures and influences that have helped create the current assessment culture in Canada and internationally; (b) the research questions are presented, with discussion of the dependent and explanatory variables to be examined; (c) some further motivations for pursuing this study are presented; and (d) the Canadian context for large-scale provincial assessments is examined.

1.1.1 The roots of accountability

The key function of standardized tests in the eyes of government is accountability. The story of today's educational testing really begins with the New Public Management (NPM) model introduced in the 1980s and its emphasis on transparency with public funds and functions (Morris, 2011; Morgan, 2009). There is little argument that information should be available to the public about schools and their effectiveness (Morris, 2011); however, what kind of information, and presented in what way? In public policy systems (in education and beyond) it is thought that the new accountability functions are fitted onto pre-existing structures. Thus the metrics of accountability are afterthoughts which may not align well with the purposes of the system or even undercut them (van Thiel & Leeuw, 2002). An 'auditable' school certainly produces data, but is it necessarily a better school for the community (Espeland & Sauder, 2007)? Propper and Wilson (2003) note that data collection does not in itself indicate improvement or adherence to the stated aims of NPM reforms. In school systems, the influence of accountability data on curriculum reform is pronounced, although this is not

always its intended purpose (Fullan, 2011; Breakspear, 2012). Møller (2008) distinguishes *political and public* accountability, which seem to drive educational assessment policies, and *professional accountability*, which relates more to teachers acting as the public would expect them to (putting students first, collaborating with colleagues, etc.) as well as being devoted to professional improvement. This is why the NPM accountability model, which tends to the political and public side of the balance, is a factor thought to drain professional autonomy from educators, and influence education policy from an efficiency-based perspective.¹ It can be dangerous, though, to assume that accountability prods will always work as expected (van Thiel & Leeuw, 2002; Propper & Wilson 2003).

1.1.2 International pressures

A related factor in the design and use of provincial tests has been the growing body of international tests and the media attention that the results garner (Fullan, 2011; Morris, 2006). Referencing OECD (the Organization for Economic Cooperation and Development) materials directly, Sahlberg notes that governments take international assessment results seriously:

Many countries are reforming their education systems to provide their citizens with knowledge and skills that enable them to engage actively in democratic societies and dynamic knowledge-based economies (Sahlberg, 2006, p.261).

The OECD has been instrumental in using its Program for International Student Assessment (PISA) as a kind of 'measuring stick' for national education systems. Some countries put more stock in these numbers than others, but it is certain from looking at Canadian ministerial documents that *all* provinces are aware and concerned with how performance in PISA tests (as well as PIRLS and TIMMS assessments) can make them appear in the eyes of a critical media (Breakspear, 2012; Uljens, 2007).² The OECD is in name and function an economic (not educational) body, yet it wields great power over decision makers from the wide-

¹ "The focus on accountability uses standards, assessment, rewards and punishment as its core drivers. It assumes that educators will respond to these prods by putting in the effort to make the necessary changes." (Fullan, 2011, p. 8)

² For example, the New Brunswick Assessment Program (NBAP) has stated goals that include administering and reporting on provincial testing *as well as* coordinating the administration of international assessments. The argument is that these are not unrelated activities and that international results inform provincial policy choices.

spread use and almost universal acceptance of these tests' validity (Volante & Ben Jafaar, 2008; Morgan, 2009).

The standardized testing model pre-dates the OECD's Program for International Student Assessment which was introduced in the year 2000, but PISA has certainly added fuel to the fire. By uniting participant countries under a common regimen of testing, PISA has produced some conformity in means and goals (Uljen, 2007; Morgan, 2009). The pressure that governments feel regarding PISA scores is 'soft' since participant countries themselves are left to decide what reforms would best address any perceived national short-falls (Martens, Kerstin, Niemann & Dennis, 2010). Not surprisingly, more student testing seems to be the universal response to PISA data. PISA is also used as a stepping off point for national or provincial curriculum reform:

The PISA and PCAP [Pan-Canadian Testing Program] assessment rankings may place Saskatchewan students at a serious disadvantage for acceptance into post-secondary education programs of study, as well as employment opportunities (Saskatchewan Ministry of Education, 2012).

A less competitive PISA ranking is the justification in this case (and others) as a reason for *more* testing, but unfortunately, the same amount of emphasis is not being placed on changes to classroom instructional practice. So testing begets testing, and the type of knowledge which is promoted for use on these assessments becomes itself a form of educational currency.³

1.1.3 Other purposes of educational assessments

Aside from these accountability functions, tests are also expected to: (a) improve student outcomes; (b) stimulate professional reflection on current practices; (c) inform and initiate professional development and system-wide reform; and (d) to rank educational systems nationally and internationally (Morris, 2011; Saskatchewan Ministry of Education, 2007; Fullan, 2011; Morgan, 2009). This is not easy for any single instrument to accomplish (Volante & Ben Jafaar, 2008). A clear and defined purpose in assessment drives not only the intended use of results, but also its design and implementation (Morris, 2011). Mehrens (1998, p.4), frames the discussion on multiple-purpose testing this way:

³ ". . . in evaluation the essential beliefs concern the credibility of the knowledge systems of the parties to the evaluation. Evaluation designs are intended to establish the credibility of the knowledge the evaluation generates." (Noblit & Eaker, 1987, p.6)

However, measurement experts have suggested for some time that "tests used primarily for curriculum advancement will look very different from those used for accountability" (Anderson, 1985, p. 24) and they will have different intended and actual impacts. Likewise, tests used for high stakes decisions (e.g. high school graduation and merit pay) are likely to have different impacts than those used for low stakes decisions (e.g. planning specific classroom interventions for individual students).

This is similar to the conclusions of Taylor, Shepard, Kinner and Rosenthal (2003) whose Colorado-based study found that teachers reacted much more positively to 'standards reform' (i.e. curriculum changes) than to the testing that went alongside of it.

1.1.4 The limits of tests

With the improvement of the educational system from the ground up predicted, expectations are high for these tests. Upon examining the large-scale assessment (LSA) model, most tests in Canada and elsewhere focus on 'core' subjects, sometimes called 'curriculum narrowing'; leaving out those courses designated as less important (Koretz, 2002; Nagy, 2000). PISA assesses only science, math, and reading (Morgan, 2009), and like tests that are modelled on it, PISA leaves out domains that are more problematic for evaluation (Luke, 2011). Even within core domains, deep understanding and breadth of knowledge, which are more difficult to assess, are left aside for simple content knowledge (Ungerleider, 2006; Smith, 1991). Nagy (2000) speaks of the trade-off that faces test developers in this context. They can choose 'testable' content which is cheaper to develop and score, or they can choose extended response items which are better indicators of curricular knowledge, but suffer from subjective scoring and are harder to equate across years and borders.⁴ The validity of the results comes into question when there is not a good alignment of the curriculum taught in classrooms and the LSA, or teachers are left to create ad hoc alignment with the assessment vehicle (Morris, 2011; Mintrop & Sunderman, 2009).

⁴ "Extended responses are more able to tap a broader range of skills and objectives, and they give better curricular signals than do multiple choice items. On the other hand, multiple choice items are more reliably scored, at a lower cost than written or performance items. They also make it easier to equate tests over time." (Nagy, 2000, p.268)

1.1.5 Changes in instruction

Research into standardized assessments has been voluminous, but rarely does it consider how well teachers adapt their instruction to the results of large-scale assessments or in preparation for them. Discussions of tests' weaknesses and strengths are common (Kohn, 2001; Volante, 2011; Scriffiny, 2008; Rhoades & Madaus, 2003) as are studies looking at how data are reported (Breakspear, 2012; Martens & Niemann, 2010). International and national testing policies have both been evaluated in terms of their ideological focus and ability to promote large-scale educational reform (Fullan, 2011; Morris, 2011; Popham, 1999). Still, the question as to how different kinds of testing regimes either encourage or discourage the use of their data by teachers has not been examined at length.⁵ There is a significant gap in the current literature and little research being done about how assessment policy and practice change instruction in Canadian schools.

1.1.6 Opportunity cost

These large-scale tests cost not only a significant amount of instructional time, but also money that might be used elsewhere in the school system.⁶ Sahlberg (2010) notes that census style testing (as advocated by Hargreaves, 2008) would deliver on most of the important accountability functions LSAs are expected to perform while also being less costly and less disruptive to regular instruction. In Canada, tests are all provincially developed and piloted before use, scoring, and sharing results. Tests are administered at the classroom level, re-directing the focus of thousands of educators for days or weeks at a time. In theory, governments should themselves be accountable for following up on the appropriate use of such costly data. Right now, and until it is proven otherwise by research data, full implementation of accountability policies at the school level is spotty at best.

As an in-service high school teacher and administrator, I have witnessed first-hand how useful some assessment data can be, but many are not used for a lack of expertise or training in the school setting. I have also seen how there is a disconnect between the intended purposes and practical uses of tests that makes

⁵ To my knowledge, there has been an examination of administrators' reactions, but in Ontario only (Volante, Cherubini & Drake, 2008) and also case studies have been examined in American urban school districts by Lachat & Smith (2009).

⁶ "Significant amounts of instructional time are spent preparing for CSAP tests. A minority of teachers statewide, between 20% and 30%, spent the few weeks before CSAP preparing their students by going over sample problems and administering practice tests." (Taylor, Shepard, Kinner and Rosenthal, 2003, p. 52)

them appear to be a burden rather than a learning opportunity for school staff. It is in the pursuit of addressing these concerns and providing usable guidelines for future assessments that I was motivated to undertake this study.

1.2 The research question

My **primary research question** is:

A) Which policies in the practice of large-scale centralized student assessment produce the most positive classroom-level use of the data?

In order to answer this question, studies from the current literature will be compared with original research to gauge the different perspectives on assessments and how they are used.

The **sub-questions** are:

B) How different are the provincial policies and practices in Canada (ministry, division, and school) related to large-scale assessment?

C) How much of teachers' practices in reacting to assessment data be traced to assessment systems and related policies set at higher jurisdictions?

D) What other factors might influence a teacher to use (or not use) assessment data?

The **dependent variable** in this study is teacher use of large-scale assessment data at the classroom level which may or may not be determined by expectations from administration, divisional staff, or ministries (even though they are not easy to separate in some cases). Clearly, only when teachers *use* the data will assessment policy translate into changes in classroom instruction. It will be examined which policies lead to data use, and which uses are positive. To define positive (as compared to negative or neutral) uses, a teachers' code of professional conduct has been employed since these norms are generally accepted by educators, ministries, and the public (while it is true that they require interpretation and the use of professional discretion).⁷ The STF Code of Professional Competence is found in Annex 1.

There are several **explanatory variables** that will have bearing on this study. First of all, the relative strengths, weaknesses and the limits of standardized tests and the results data will be examined. This will be an analysis of what tests evaluate (and in what ways), as well as how the data are presented and how they are interpreted (and by whom). These factors are expected to vary between provinces, and translate into different forms of reactivity. Test designs obviously favour selected learning and teaching (content) above that which is not selected

⁷ "For a [professional teacher] code to be considered effective, it must be framed for the membership to influence positive behaviours." (Nuland & Poisson, 2009) This criterion is applied in terms of the reactivity model, as well.

(Noblit & Eaker, 1987; Shepard, Davidson & Bowman, 2011; Yeh, 2001). Research has shown that teachers' view of test design and content matters to implementation (Ungerleider, 2003; Volante, Cherubini & Drake, 2008; Ryan & Joong, 2005).

Second, the role of policies that support or create incentives will be examined. This variable will include observations on teacher professional development (PD) and collaboration related to testing and data, as well as those policies which set explicit expectations for the use of assessment data (commonly called 'stakes'). Again, provinces have different ways of carrying out large-scale assessments, and reactions should differ along with these policies. If teachers are to use the data, these must be reported in such a way that renders them usable (Young, 2006; Volante & Ben Jafaar, 2008; Halverson, 2010). PD directly related to assessment is often required and requested by teachers to make LSA policy functional (Scott, Webber, Aitken & Lupart, 2011; Kemp & Freisen, 2009; Schorr, Firestone & Monfils, 2003).

Finally, the role of the teacher in reaction to large-scale assessment will be examined. The relative effectiveness of policy is often determined at the classroom level since test design and policy alone cannot dictate how effective implementation will be.⁸ Interviews will determine in what ways the actors at the most basic level of authority in schools (i.e. teachers) use their professional discretion to interpret policy. This aspect of the study will call upon current literature on policy implementation and how it may be resisted, interpreted, and altered by 'street-level' bureaucrats (Pressman & Wildavsky, 1973; Alexander & Faludi, 1989; Elmore, 1980; Matland, 1995). The skills and experience each teacher brings to the table will also be examined for influence on data use.

1.3 Motivation

Instructional change at the classroom level, the most basic unit of public schooling, is all dependent on the instructional leader at that level, the teacher. Little work has been done to examine the interplay of jurisdictions in Canadian testing policy according to Volante, Cherubini & Drake (2008)⁹. An often

⁸ Cimbricz (2002, p.14) states: ". . . while state testing does matter and influence what teachers say and do, so, too, do other things, such as teachers' knowledge of subject matter, their approaches to teaching, their views of learning, and the amalgam of experience and status they possess in the school organization. . . the influence state-mandated tests has (or not) on teachers and teaching would seem to depend on how teachers interpret state testing."

⁹ "Little attention has been directed at understanding leadership with student assessment and evaluation as it is effected by district and provincial level factors." (Volante, Cherubini & Drake, 2008. p. 7)

overlooked but important fact is that when international comparisons of test scores place 'Canada' as a single entity, these neglect to note that as a nation it has ten different education policy jurisdictions (Woessman, 2001).¹⁰ Canada does not have a uniform education system, and does not even have a national education ministry. By examining the role of front-line staff related to administering tests and analyzing standardized assessment data in each of Canada's ten provinces, this study intends to reveal which practices best facilitate teacher preparedness and effectiveness in making sound instructional decisions based on the data. These different assessment models should be distinguishable by the practices of actors in lower jurisdictional levels. Note how Darling-Hammond and Rustique-Forrester (2005, p. 311) list the various influences upon the effectiveness of testing:

Among the factors that appear to influence the outcomes of testing are the nature of tests (what kinds of things are assessed and how); the uses of tests (what kinds of decisions are made based on test scores); the capacity for improvement represented by teacher knowledge and skills; and the context for school improvement at the state, district, and school levels, including resource levels and professional development opportunities.

Whereas test design and its stated purposes are provincial decisions, teacher improvement is often a function of division or school level policy, and both resources and professional development are most often determined by local (not provincial) needs. This study will attempt to account for these differences and similarities across the nationwide data set.

As policies are made and enforced at different levels, Ravitsch (2010) makes the important point that politicians (meaning, in that case, state officials), are not necessarily professional educators, and yet they feel able and inclined to make decisions that affect schools, teaching, and students in significant ways. Corcoran and Goertz (1995) note that it is often difficult for jurisdictional hierarchies to coordinate their work since they have different concerns, different abilities, and different mandates.¹¹ Yet this type of coordination is vital to the

¹⁰ Woessman in particular misses the wide variation in provincial policies: "This is because the institutions within a country do not vary enough to test how different institutions affect student achievement. Only the international evidence, which encompasses many education systems with a wide variety of institutional structures, has the potential to show which institutions heavily affect student performance." (p. 68)

¹¹ "States have a particularly difficult time linking teacher policies to their standards and accountability policies. Districts seem to have difficulty linking personnel policies and professional development to standards. Schools find it easier to create structures for

success of large-scale assessment programs, to improve outcomes at the school-level and also to broad-based educational reforms (Volante & Cherubini, 2007). Dorn (1998) argues that investment in professional development and in-services for (especially) new teachers are effective means to support increased assessment literacy so that educators can understand and utilize LSA data. Teachers sometimes see in this 'disconnect' (the insufficient professional support for LSA testing systems) a motive to de-professionalize them and enforce a 'cookie cutter' model for all schools that dismisses educators' unique insights into their classrooms and communities (Zigo, 2001).

There is little doubt that standardized assessment in one form or another is here to stay in Canadian public education. The question becomes, then, how it can best be designed and supported in order to improve the quality of instruction in classrooms. If this question were the primary focus of accountability testing, and if it were done using proven methods of policy implementation, then regardless of improved scores, it is likely provinces would have stronger, more rigorous educational systems as a result of having better teachers striving for professional growth within them.

1.4 Canadian context

1.4.1 National demographic and educational information

Canada is a nation with a population of 35 427 542.¹² Confederated in 1867, it was a British colony until becoming an independent nation. The British North America Act of 1867 created the new country, and the constitution was repatriated by the Constitution Act of 1982. The nation has always divided powers between the federal government and the provincial governments. Education, the focus of this dissertation, has always been a provincial responsibility. Canada began with 4 provinces in 1867 and has expanded to include 10, the last joining in 1949 (Newfoundland and Labrador). Three partially self-governing territories exist in the north, but they do not share the powers given to provinces by the Constitution.

There is no national Ministry of Education since the provinces are in control of this aspect of public service. The closest thing to a national body is the awkwardly named Council of Ministers of Education, Canada (CMEC).¹³ Some

collaboration than they do developing effective focusing mechanisms." (Corcoran & Goertz, 1995, p. 30)

¹² As of Q2, 2014, from Statistics Canada. Census data retrieved Aug. 9, 2014 from: <http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0510005&paSer=&pattern=&stByVal=1&p1=1&p2=31&tabMode=dataTable&csid=>

¹³ CMEC information retrieved Aug. 9, 2014 from: <http://www.cmec.ca/>

agreement between provinces regarding curriculum has yet been accomplished (these are regionally constructed – the Western and Northern Canadian Protocol [WNCP, 1993]¹⁴ and the Atlantic Provinces Education Foundation [APEF, 1995]¹⁵ are examples).

There are no federal curricula or federal schools with some important caveats to these generalizations. All First Nations peoples were guaranteed a federally-funded education by the Indian Act of 1876. In modern times, First Nations students have their schooling funded by the federal government (using a much less generous funding formula than the provinces), but schools are generally run very much like provincial schools. Canada's aboriginal population is distributed in all the provinces and territories with the largest populations in Ontario and provinces west of there (Manitoba, Saskatchewan, Alberta and British Columbia). As a proportion of provincial populations, though, Manitoba and Saskatchewan have the highest numbers of First Nations peoples (see **Figure 1.3** below).

The children of Canadian military service-people stationed overseas are also exceptions to the stricture on federal schools. These men and women are federal government employees and also have the opportunity to send their dependents to schools funded and overseen by the federal government. Once again, though, they are run much like provincial schools following the curriculum of a chosen jurisdiction (SHAPE International School in Belgium and AFNORTH International School in Germany both follow the guidelines of the Department of National Defence's "Schools Overseas" program and offer courses based on the Ontario curriculum).¹⁶

Canada is a nation that has two official languages with the majority of Canadians (79.4%) acknowledging that English (of the two official choices) is their primary language whereas 20.6% indicate French is their primary tongue. The French-speaking population is centered in Québec and the largest population of Francophones outside Québec lives in New Brunswick. Public schools instruct in either English or French, and even though for this study only English language schools were considered, there are 6.5 million Canadians for whom French or English is not their first language.¹⁷

¹⁴ WNCP information retrieved Aug. 9, 2014 from:

<https://www.wncp.ca/english/wncphome.aspx>

¹⁵ APEF information retrieved Aug. 9, 2014 from:

http://www.ednet.ns.ca/files/reports/essential_grad_learnings.pdf

¹⁶ Department of National Defence education information retrieved Aug. 9, 2014 from:

<http://www.forces.gc.ca/en/caf-community-support-services/dependent-education.page>

¹⁷ Statistics Canada, retrieved Aug 9, 2014 from: <http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo11a-eng.htm>

of provincial assessment, although the grade levels, the subjects tested, and the types of test (for accountability or transparency, for diagnostic purposes, to set minimum standards for graduation, or diploma exams which set common standards for high school graduates) differ (see **Figures 1.1 and 1.2**).

In examining the stated goals of provincial tests, some significant differences appear across the provinces. These will be further examined as each of the provinces and their testing policies are described. Even though there is significant cooperation on curriculum goals and text resources (see WNCP and APEF above), testing policy does not seem to have the commonalities that are evident in these other aspects.

Figure 1.2: Purposes of provincial testing in Canada

	Policy level		Classroom level					Common to all		
AB	X	†		X	X			X	X	6
BC	*	†	*		X	*	X	X	X	8
MB	X	†			X		X	X	X	6
NB	X	X	X	X	X	X		X	X	8
NL	X	†		X		X		X	X	6
NS	X	†	X	X		X	X	X	X	8
ON	X	X	X	X	X	X	*	X	X	9
PEI			*	X	X			X	X	5
QC	X	X	X	X	X	X		X	X	8
SK	*	‡	*		*	*		X	*	7

X Purpose evident in ministry literature
 * Not explicitly stated, but apparent from ministry literature
 † Exams must be written but need not have a passing grade
 ‡ Exams are mandatory when teachers are non-accredited

Seven provinces have graduation requirement exams, but in different subjects, and at different grade levels (for one of these seven, the exam is only written if the teacher is not accredited in that subject). All provinces but two wait until at least grade 3 to start LSAs. One province gives no LSAs from grades 10-12.

<http://www.statcan.gc.ca/pub/81-582-x/2013001/tbl/tblc2.2-eng.htm>

<http://www.statcan.gc.ca/pub/81-582-x/2013001/tbl/tblc2.3-eng.htm>

This kind of variation makes it somewhat difficult to generalize about large-scale provincial testing in Canada.

As opposed to subjects and grade levels, the purposes for which testing is done are relatively similar across the ten provincial jurisdictions (**Figure 1.2**). In fact, there were two points of reference common to all ten assessment policies as spelled out on ministerial websites. These websites were the starting point of this chart which identifies the ten purposes made most explicit by assessment policy makers. Clearly there is a lot expected of these assessment instruments – anywhere from five to nine purposes were made explicit or understood from education ministry information.

1.4.2 Provincial demographic and educational information

Alberta

"The purpose of the Achievement Testing Program is to: determine if students are learning what they are expected to learn; report to Albertans how well students have achieved provincial standards at given points in their schooling; assist schools, authorities, and the province in monitoring and improving student learning . . . Careful examination and interpretation of the Achievement Testing Program results can help reveal areas of relative strength and weakness in student achievement."²²

Alberta is a prosperous province in western Canada. Running from the Rocky Mountain foothills to the mid-continent prairie, it has many natural wonders for the traveler. Thanks to abundant resource wealth, it remains the only province in Canada that does not charge provincial sales taxes. The wealth of the province is also evident in the fact that its proportion of the national GDP is larger than Alberta's proportion of the national population.

In terms of provincial assessment, Alberta tests all students in grade 3, 6 and 9 in core subjects (English and math in grade 3, these as well as science and social studies in grades 6 and 9). The diploma exams of grade 12 are not a graduation requirement, but two grade 12 exams must be written to graduate and are assigned a fixed value of 50% of a student's final grade in those subjects. Students must write English and social studies tests, but math and science are dependent on course selections.

²² Alberta Education, retrieved Aug. 9, 2014 from:
<http://education.alberta.ca/admin/testing/achievement-results.aspx>

According to the 'testing results' page at Alberta Education, the LSAs are explicitly intended to make teachers reactive and deal with the data in constructive ways. To improve teaching and learning *is* reactivity, but how exactly this policy is implemented in classrooms remains to be determined. The diploma exams are commonly referenced as a 'common standard' to help remedy mark inflation at the school level. Alberta Education makes reference to six of the nine 'purposes of assessment' (compiled in **Figure 1.2**) as objectives for their assessment policy. The focus seems to be more on policy-level issues than classroom-level ones - interventions for struggling students being the only one of the latter mentioned.

British Colombia

“Using information from FSA, the Ministry of Education works with school districts to provide support for students and to improve teaching and learning for the coming school year. . . The BC Performance Standards are intended as a resource to support ongoing instruction and assessment.”²³

British Columbia is Canada's westernmost province with a long, rugged, and beautiful coastline running along the Pacific from Washington State to the Alaskan panhandle. The province's topography is dominated by the Rocky Mountains which run down the length of BC. It has a slightly larger proportion of the Canadian population than Alberta, but the numbers of enrolled students (in the 2011 school year) were almost identical (a difference of 19 students). It also has an abundance of resources and a diverse economy from saw mills in the interior to manufacturing and technology sectors in the Vancouver area.

In terms of LSAs, British Columbia has a fairly standard program of testing – English and math in grades 4 and 7 (called 'Foundation Skills Assessments' or FSAs), and a richer mix of exams called 'provincial exams' (including grade 10 Science, grade 11 Social Studies and Civics, and grade 12 Communications and BC First Nations Studies) the writing of which are graduation requirements and the marks are blended with classroom grades at the high school level (20% of final marks in grades 10 and 11, 40% in grade 12 except for BC First Nations Studies 12, which remains 20%). Writing these exams is a requirement for graduation, but it is not a requirement to pass the exam. The objectives of the assessment policy are many – eight of the nine identified by the research across all ministry websites (**Figure 1.2**). The province has justified the time and the expense of the assessments

²³ British Columbia Education, retrieved Aug. 9, 2014 from:
http://www.bced.gov.bc.ca/assessment/fsa/pdfs/fsabrochure_print.pdf

in trying to make parents aware of the program.²⁴ Interestingly, in BC, as compared to other provinces, the level of controversy the assessments create and the amount of resistance these exams meet from the public and the teachers' union (the BC Teachers' Federation) is considerable.

So while it is apparent from the ministry that the exams are intended to be used for improved instruction ('a resource to support. . .') and to guide teacher's decisions with solid data, the controversies surrounding the assessments currently undercuts these classroom-level intentions quite severely. There has been labour unrest in the BC education sector as well, with a strike being called at the end of the 2014 school year. This does not improve the chances that teachers will take seriously the ministry guidelines to use these data for instructional purposes.

Manitoba

"The Provincial Assessment Program supports learning by: providing feedback to students, teachers and parents about student learning; informing instructional planning and helping to determine the need for changes or student specific interventions; providing system-wide information that assists in identifying trends and making decisions about resources and support; providing the public with general information about student achievement to sustain confidence in the education system"²⁵

Manitoba is the province located at the geographical heart of Canada, and it is in many ways a bridge between the east and west of the nation. Canada's first rebellion started here at the Red River Colony in 1869 and was fought partly over the rights of citizens (Métis citizens in this case) to a public education in their language and their faith. Manitoba's tradition as a place of exchange and trade was built on the backs of trappers, voyageurs, and the Canadian Pacific Railway which, once completed, went straight through to the Pacific Ocean. This rich history of settlement and First Nations is still evident in current demographics, as well (see **Figure 1.3**).

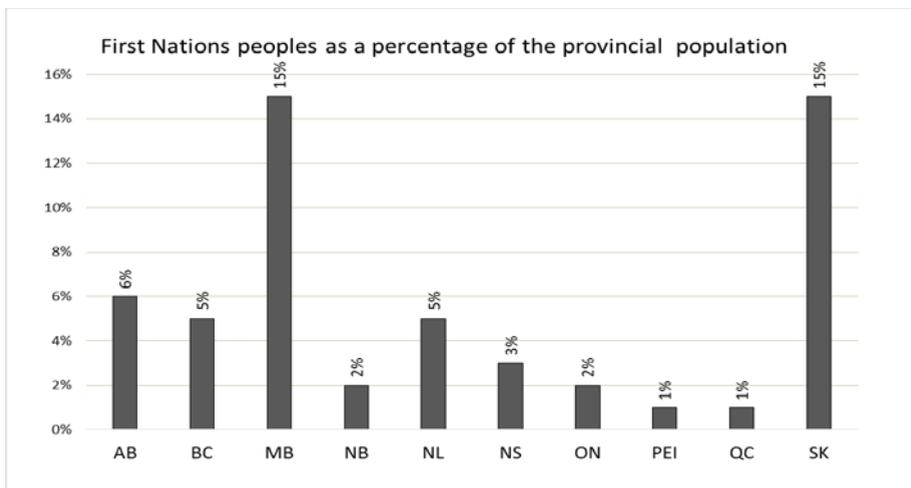
Manitoba Education has the most unique system of provincial assessment in the nation. Both English and math assessments are done in grade 3, math again

²⁴ BC Education: "The entire assessment takes about four and a half hours to complete, and most schools spread them over the course of two or three days. That's less than 10 hours in total of provincial assessment from Kindergarten through Grade 9. . . FSA is not expensive. It costs about \$20 for each student in Grade 4 and Grade 7, or an average of \$2 a year over a student's first 10 years of schooling."

²⁵ Manitoba Education, retrieved Aug. 9, 2014 from: <http://www.edu.gov.mb.ca/k12/assess/>

in grade 7, and English again only in grade 8. Yet these assessments are not paper-and-pencil tests written by the students. They are outcome-based checklists that are completed by the classroom teacher over the course of several months. In this way, the assessments are 100% aligned with the curriculum. They also put no test pressure on the students and for them require no special instruction or preparation.

Figure 1.3: Manitoba and Saskatchewan have the largest number of First Nations peoples as a proportion of their total population.²⁶ First Nations students have much lower academic success rates than non-aboriginal students, and this is an ongoing challenge in these (and other) provinces.²⁷ Compared to other provinces, Manitoba has historically had the lowest overall high school graduation rate which is at least partly attributable to the low graduation rate and high population representation of First Nations students.²⁸



Negative reactivity effects (commonly known as 'teaching to the test') would be hard to trace on this type of assessment. On the other hand, they could be informative enough (by guiding a teacher to establish for certain which specific

²⁶ Population data from Statistics Canada, retrieved Aug 9, 2014 from: <http://www.statcan.gc.ca/pub/89-645-x/2010001/c-g/c-g004-eng.htm>

²⁷ According to the Assembly of First Nations, aboriginal graduation rates (at 36%) are only half of those for non-aboriginal students (72%). Retrieved Aug. 9, 2014 from: http://www.afn.ca/uploads/files/events/fact_sheet-ccoe-3.pdf

²⁸ Graduation rate data from Statistics Canada, retrieved Aug 11, 2014 from: <http://www.statcan.gc.ca/pub/81-595-m/2011095/tbl/tbla.11-eng.htm>

outcomes are not being met by which students) to create positive reactivity effects (more on reactivity follows in Chapter 2).

At the high school level, the assessments are more in line with what is seen in other jurisdictions. There are four mandatory provincial tests, all at the 40 level (grade 12): (a) Applied Mathematics; (b) Essential Mathematics; (c) Pre-Calculus Mathematics; and (d) English Language Arts. These tests are assigned a set proportion of the student's mark for the final class grade (30% of the final mark for all exams except Essential Math, which is 20%). Writing exams in these classes is a requirement for graduation, but they need not necessarily be passed. Manitoba Education expects these test results to be used for public accountability, but also for classroom interventions, and benchmarking to international standards set by PISA.

New Brunswick

“Assessment enables teachers to gather data to determine the needs of their students, and to address those needs adequately in order to tailor instruction. Large-scale data gathered through provincial assessment programs enables policy makers to make programming decisions at the provincial, district or school level.”²⁹

New Brunswick is one of Canada's five Atlantic Provinces and features the Bay of Fundy on its eastern shore, rivers all along the coast, and it borders the Gaspé region near the Baie des Chaleurs to the north. New Brunswick has the nation's second highest rate of French-speakers (by population) and lacks the animosity or tribalism that seems to taint Québec language politics. It is accepted in New Brunswick that English- and French-speakers will both find opportunity and welcome everywhere (see demographic detail in **Figure 1.4**).

Student assessment in New Brunswick is quite comprehensive, including tests in grades 2, 3, 4, 5, 7, 8, and 9 (this is a similar amount of testing as other provinces but not here concentrated at specific grade levels). The framework document (cited below) explains in great detail not only the tests, but the rationale, and their intended purposes. Page 4 of that document goes so far as to spell out chapter and verse from the provincial Education Act the roles and responsibilities of teachers, school districts, the Minister's advisory committee (called the Minister's Advisor Council on Testing and Evaluation – MACTE), the Minister him

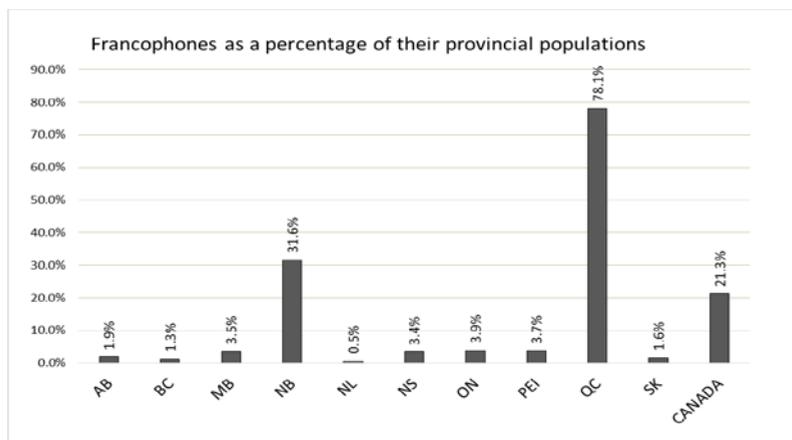
²⁹ New Brunswick Department of Education and Early Childhood Development, retrieved Aug. 9, 2014 from:

<http://www.gnb.ca/0000/results/pdf/AssessmentFrameworkDocument.pdf>

or herself, the Assessment and Evaluation Branch and the Department of Education. Clearly provincial assessment is a joint venture in New Brunswick.

The tests themselves are focused only on English (or French) and math skills. The grade 9 English test is called the English Language Proficiency Assessment (ELPA) and is a minimum competency exam that must be passed to graduate high school, and is specifically linked to international assessment standards: “This aligns with the [OECD] definition of functional literacy.”³⁰ It can be re-written in grades 10 or 11 as the ELPR ('R' for 'reassessment') if it had not been written or passed before this time. The department also outlines eight objectives for their assessment policy including policy-level and classroom-level purposes for the data.

Figure 1.4: Next to Québec, which is primarily French-speaking, New Brunswick has the largest proportion of its population who identify French as their mother tongue.³¹



Newfoundland and Labrador

The information obtained from these assessments assist in improving student achievement, evaluate the effectiveness of provincial programs, inform parents and students of performance based on curriculum outcomes, and set expectations of what

³⁰ New Brunswick Department of Education and Early Childhood Development, p.10. Retrieved Aug. 9, 2014 from:

<http://www.gnb.ca/0000/results/pdf/AssessmentFrameworkDocument.pdf>

³¹ Linguistic information by Statistics Canada, retrieved Aug. 9, 2014 from:

<http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo11a-eng.htm>

students should know and be able to do by the end of Grades 3, 6, and 9.”³²

Newfoundland and Labrador is Canada's most recent provincial addition since joining Confederation in 1949. It was likely the first place in North America visited and colonized by Europeans (at L'Anse aux Meadows), and was settled very early on especially because of its abundant resources of fish. The island is well known in Canada for the hospitality of its people, their unique linguistic traditions (there is an official Dictionary of Newfoundland English), and having such a rugged terrain that it is nationally known as “the Rock.”

Provincial assessment in Newfoundland and Labrador follows similar patterns to those seen in other jurisdictions, but is not exactly the same. Testing in English and math is done at key levels (grade 3, 6, and 9), and public exams, as they are known, are written in grade 12 in most academic subjects (English, French, math, several sciences, and a few social sciences). As the Department of Education website makes clear, “Exams are marked by an independent marking board made up of teachers. The final mark in each of these courses is 50 per cent school mark, and 50 per cent exam marks.”³³ These tests tend to be some of the more high stakes exams in Canada since they count for so much of the final grade. Student pressure often translates into teacher pressure as well, so grade 12 teachers in this province are well aware of the requirements of these exams.

Six of the nine purposes of assessment (see **Figure 1.2**) are expected from these provincial assessments with a heavy focus on policy-level purposes: (a) accountability; (b) data for central decision making; (c) adherence to curriculum; (d) the monitoring achievement; and (e) benchmarking achievement related to international assessments.

Nova Scotia

“The objectives of Evaluation Services are to: administer the Program of Learning Assessment for Nova Scotia (PLANS), which includes provincial, national and international assessments, in French and in English; develop and administer program assessments to determine the effectiveness of curriculum delivery; develop and administer student assessments to assist students to achieve outcomes; provide student achievement information to

³² Newfoundland and Labrador Department of Education, retrieved August 10, 2014 from: <http://www.ed.gov.nl.ca/edu/k12/evaluation/crts/index.html>

³³ Newfoundland and Labrador Department of Education, retrieved August 10, 2014 from: <http://www.ed.gov.nl.ca/edu/k12/evaluation/exams.html>

government for education decision making; help teachers understand assessment principles and practices; support school accreditation through collecting, analysing and reporting results of questionnaires, which helps to improve education decision making; publish accountability reports for all assessments and examinations, both for teachers and for the general public.”³⁴

Nova Scotia is an Atlantic Province which has some spectacular historic sites and seascapes. From Cape Chignecto where the Fundy tides roll into the Minas Basin, to the highlands of Cape Breton along the Cabot Trail, and into historic Halifax where the landscape is dominated by the former British fort on Citadel Hill and the expansive Commons, Nova Scotia is rich in scenery. It has always been tied economically to the waters that surround it, be it with fishing, ship-building, or trading. Halifax is one of the largest and deepest natural open-water ports in the world and it sees as a result a considerable amount of container ship traffic.

In terms of provincial assessment, Nova Scotia has just embarked on a new testing program that started in the 2013-2014 school year. Some of the questions asked on the research questionnaire were just a few months too early for them to answer. That being said, there had been provincial assessments in place prior to the last school year, so a picture of the assessment program in general was possible to generate based on survey and interview responses. Assessments are given in grades 3 (English) and 4 (math), grade 6 (English and math), and grade 8 (English and math). Exams (note that a different term now applies) for grade 10 English and Math are assigned 20% of the student's final grade in the course. All students, excluding those on individualized learning plans or with special circumstances, must write this exam. This exam is marked locally by the class teacher, and in the following summer provincial teams re-mark the exams as a means of passing information on to policy makers. This allows for both classroom-level and policy-level choices to be scrutinized.

Policy makers expect the assessments and exams to be used for many purposes (eight of the nine identified by the researcher – see **Figure 1.2**) so this 'dual-marking' system may be a means of providing different information to different stake-holders. Teachers, parents and students want scores, while school boards and Department of Education officials want to verify curriculum adherence, accountability and more data with which to inform policy choices.

³⁴ Nova Scotia Education and Early Childhood Development, retrieved Aug. 10, 2014 from : <http://plans.ednet.ns.ca/about-plans>

Ontario

“EQAO's tests measure student achievement in reading, writing and mathematics in relation to Ontario Curriculum expectations. . . By providing this important evidence about learning, EQAO acts as a catalyst for increasing the success of Ontario students. The objective and reliable results from EQAO's tests complement the information obtained from classroom and other assessments to provide students, parents, teachers and administrators with a clear and comprehensive picture of student achievement and a basis for targeted improvement planning at the individual, school, school board and provincial levels. EQAO helps build capacity for the appropriate use of data by providing resources that educators, parents, policy-makers and others in the education community can use to improve learning and teaching.”³⁵

Ontario is Canada's most populous province and is also the second largest by area. It covers ground from the Great Lakes in the south, to the national capital of Ottawa in the east, far into the north touching on Hudson Bay, and a far west as to be almost in the geographic centre of the continent. Ontario has both wilderness and urban life. Its very size and prominence make Ontario the subject of some derision by non-Ontarians, and until the Toronto ice hockey team (the Maple Leafs) wins a Stanley Cup, there will always be fodder for this antipathy.

Ontario's Education Quality and Accountability Office was established in 1996 and is different from the assessment branches found in most provincial ministries in that it is not found in the ministry. It is designed as an 'arm's length' institution (technically a 'Crown agency') to provide reliable information to Ontario students, citizens and politicians about educational achievement. They share an extensive amount of material on assessment (certainly not just results) with the public and with politicians, and in this respect their mandate seems to have been well-managed.

The early years testing done in Ontario is not significantly different from most other jurisdictions as the grade levels and subject matter is much the same (grade 3: English and math; grade 6: English and math; grade 9: math; and grade 10: English). Only this last exam, the so-called OSSLT (Ontario Secondary School Literacy Test) is of a different sort. There is only one other province that has a minimum competency graduation exam that must be passed to graduate (New

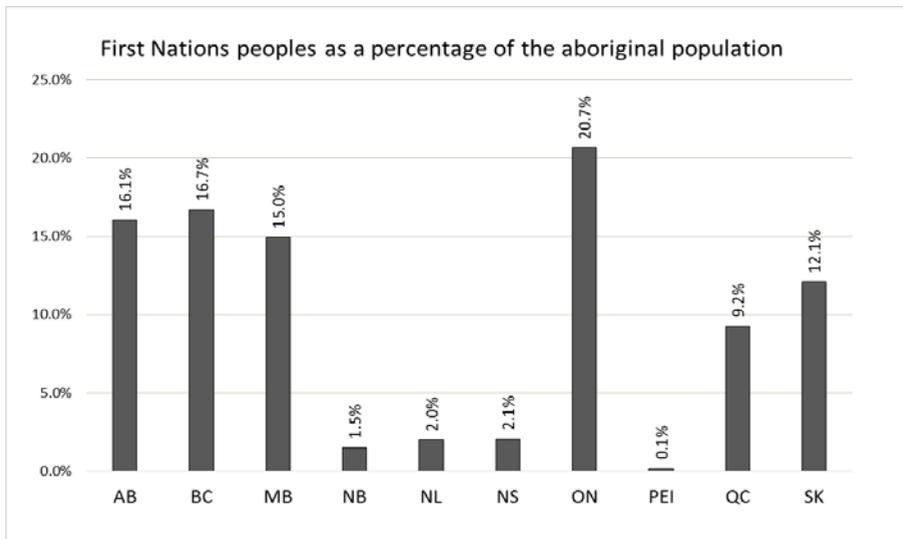
³⁵ Education Quality and Accountability Office, retrieved Aug. 10, 2014 from: <http://www.eqao.com/AboutEQAO/AboutEQAO.aspx?Lang=E>

Brunswick). This certainly focuses the attention of students and teachers alike on this exam. It can be rewritten in future years if not passed in grade 10.

Ontario has a largest number of public school students of any province, and also the largest number of First Nations students. Given the size of the entire student population these aboriginal students are not a highly significant part of the student population. **Figures 1.3** (above) and **Figure 1.5** (below) shows this contrast – a high number of First Nations students mixed in with the largest provincial student population in Canada.

The mandate of EQAO is to share assessment information for many purposes, yet it remains true that the OSSLT is one of least reactive of the assessments (since once passed, it is not re-visited by teachers at all).

Figure 1.5: One in five First Nations people live in Ontario (where they constitute just 2% of the provincial population)³⁶



The grades 3, 6 and 9 assessments, on the other hand, could be used (and according to EQAO literature, should and must be used) to inform instruction. This being the case, the researcher was surprised to get a rejection letter from one school board in Ontario to which an application had been made to distribute my teacher survey. The email stated:

³⁶ Population data from Statistics Canada, retrieved Aug 9, 2014 from: <http://www.statcan.gc.ca/pub/89-645-x/2010001/c-g/c-g004-eng.htm>

“We are not able to accommodate your request at this time, as your study does not align with the practices of the . . . School Board with respect to the use of EQAO data. . . Please note that the Board uses EQAO data at the district and school levels in conjunction with other data sets. . . The Board does not support the use of EQAO data at the individual teacher level which is the primary focus of your study.”

It seems the EQAO's stated mission to provide information to individuals and teachers for targeted improvement has not been made clear to leaders in this Board. It is not surprising that there was some re-interpretation of EQAO's message in this way, but it does indicate that different Boards have very different expectations for the provincial test data. It should be expected that without instructional change, there is no way to address the EQAO scores at all. To speculate somewhat, teachers likely actually *are* reactive to these data, but the Board is either unaware of or unsupportive of the practices. All of the nine purposes of assessment (compiled in **Figure 1.2** from ministry websites) are expected from EQAO assessments.

Prince Edward Island

“Provincial assessments are conducted yearly and tell us how well students are doing at key stages of learning. Students are assessed in reading, writing and mathematics at the end of Grade 3, Grade 6, and Grade 9. We also have an assessment called Early Years Evaluation (EYE) that takes place before a child goes into kindergarten. These assessments are developed by teachers from across the province and are based on the curriculum used in Island schools. They tell us how well students are learning the curriculum, where students may need help, and how resources may be directed to improve our education system.”³⁷

Prince Edward Island is Canada's smallest province by area and population. It is also very scenic and represents some of Canada's best known agricultural (potatoes) and cultural (the books of Lucy Maud Montgomery and the iconic music of Stompin' Tom Connors) exports. The Confederation Bridge, which was an architectural first at the time it was built in 1993, connects the island to New Brunswick and the name is a tip of the hat to Canada's founding Confederation document, signed in Charlottetown, PEI on July 1, 1867.

³⁷ Prince Edward Island Department of Education and Early Childhood Development, retrieved Aug. 20, 2014 from: <http://www.gov.pe.ca/eecd/studentassessment>

Provincial assessment in PEI began in 2007 but is the most limited in scope of all the provinces. Students in grades 3, 6, and 9 write English (or French) and math assessments. No exams at all are written in high school, so no graduation requirements are based on external assessments. Many economists have declared that external exits exams (which is the function of many high school exams in other Canadian jurisdictions) are the best way forward in educational attainment (Bishop, 1998; Wößmann, 2003), but the evidence from the graduation rate of PEI students (from 2005-2008, first in Canada, in 2008-2009, second, 2009-2010 fifth),³⁸ and high school attainment rates over the last 15 years (from 1997-2012 it has increased by 16.5% ³⁹ which is the highest rate of improvement in Canada) seem to point in another direction.

The scope of the purposes to which assessments will be put is also limited in PEI. Only five of the nine purposes flagged by the researcher (in **Figure 1.2**) are evident in PEI provincial assessment policy, and the policy-level goals are limited to curriculum adherence and collection of data for decision-making purposes. The sole classroom-level goal is to use the data for interventions to assist struggling students.

Québec

“The intention of this examination is to provide opportunity for students to demonstrate knowledge and competency, as well as to provide teachers the opportunity to judge literacy development. Data obtained from student performances on the tasks prescribed in this examination, in conjunction with data collected from performances during the cycle, will help the teacher form judgments about the levels of competency attained by the end of the cycle for the end-of-cycle report.”⁴⁰

Québec is a vibrant and politically charged part of Canada. It is the largest province by area, and the second largest by population. It is mostly French-speaking (see **Figure 1.6**) and has been fairly adamant that this remain so despite the incursions of English into Québécois culture. Bill 101 is the most striking

³⁸ Graduation data from Statistics Canada, retrieved Aug 9, 2014 from: <http://www.statcan.gc.ca/pub/81-595-m/2011095/tbl/tbla.11-eng.htm>

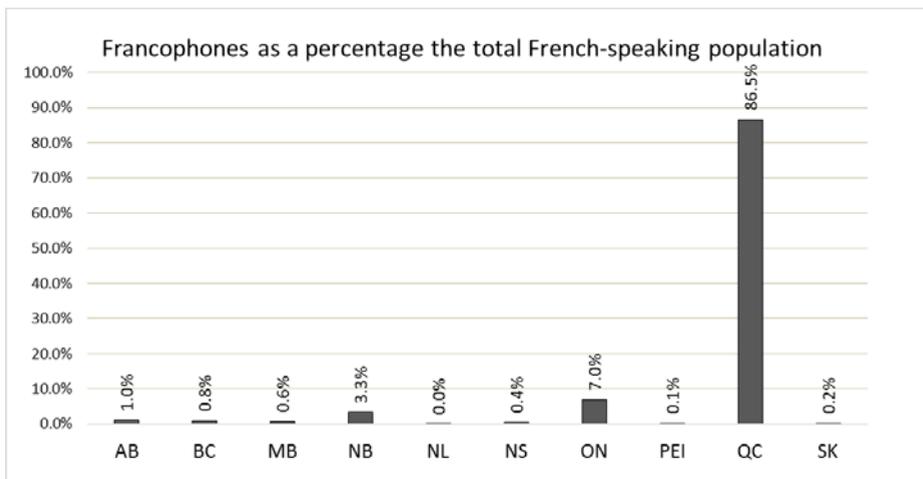
³⁹ From the Conference Board of Canada, retrieved Aug. 11, 2014 from: <http://www.conferenceboard.ca/hcp/provincial/education/highschool.aspx>

⁴⁰ Québec Ministère de l'Éducation, du Loisir et du Sport, retrieved Aug. 25, 2014 from: http://www.learnquebec.ca/export/sites/learn/en/content/curriculum/languages/ela/documents/InfoDoc_ELA_Cycle3Prim.pdf

example of this – a law mandating the use of French in Québec society so strict that in 1999 Old Navy (the retail chain) was officially asked to change its name in Québec locations to 'La Vieille Rivière' (they did not comply). Even so, the cultural richness of Québec, with the Winter Carnival, the jazz and comedy festivals, and the beautiful architecture of the French-colonial era make it a unique province.

Québec (like New Brunswick to a smaller scale) has parallel and equal English and French school systems and assessment policies. High school testing in Québec is done in several subjects over the final two years of secondary cycle 2 (what would be grades 10 and 11 in other provinces). In grade 10, math, social studies and science exams are written. In grade 11, English and core French exams are written. These exams make up a fixed portion of a student's final grade. However a unique function known 'moderation' is used in Québec by the ministry during the marking process. If a student's exam result is much lower than the classroom grade, the ministry can *moderate* this effect, essentially decreasing the difference by lowering the classroom mark in the pursuit of a more accurate reflection of the student achievement. This is the only province that has the ability to change teachers' classroom grades beyond the more common blended marks which include both classroom grades and provincial exam scores.

Figure 1.6: Most French-speakers in Canada live in Québec, "la belle province", followed far behind by Ontario and then New Brunswick.⁴¹



LSAs are also written prior to high school in English (or French) and math at the end of Cycle 1. Students tend to be moving on from their elementary schools

⁴¹ Linguistic information by Statistics Canada, retrieved Aug. 9, 2014 from: <http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo11a-eng.htm>

directly after writing these tests, so the data would need to be shared between schools and teachers to be of use to the instruction of the students who write the tests. Of course, it is possible for teachers in Cycle 2 to make use of the data to amend their instruction, but this practice is less common than policy makers would prefer.

Saskatchewan

“The *Help Me Tell My Story* assessment includes a variety of unique pre- and post-assessment supporting material to ensure children are comfortable and familiar with the assessment. . . The holistic assessment is comprehensive, ensuring it goes beyond measuring simply proficiency in oral language. . . Finally, the assessment results are immediately linked to activities, or learning ideas, accessible to all teachers, caregivers and Elders. The connection to activities help ensure that the data leads to immediate action.”⁴²

Saskatchewan is the least populous western province, and is a map-makers dream with man-made boundaries on all sides. The joke goes, 'Saskatchewan - easy to draw, tough to spell.' It leads Canada's agricultural production of such crops as wheat, oats, and lentils.⁴³ It is also a province with surprising geographical diversity from the numerous beautiful lakes of the north and the Canadian Shield, the grasslands of the southwest, surprising sand dunes around Lake Athabasca, and the sole mountain-level elevation at Cypress Hills. Do not take the name of the Saskatchewan Roughriders in vain here, though – the friendly residents are immensely protective of their proud Canadian Football League legacy (yes, there is such a thing as *Canadian* football).

Provincial assessment in Saskatchewan is currently in a period of transition, and this made data-collection in this province less straight-forward than some others. There are new tests being piloted in 2013-2014 and into 2014-2015 for early years classes (grades 1, 2 and 3 English and math), and the older course of provincial assessments called AfLs (Assessment for Learning) had been postponed for 2 years running. As of summer 2014, all reference to AfLs has been removed from ministry websites, so it appears they are no longer part of provincial assessment policy. It is widely speculated that they are in the process of being

⁴² This refers to an English assessment, but the new Math assessment has the same general outline. Saskatchewan Ministry of Education, retrieved Aug. 11, 2014 from:

<https://holisticassessment.gov.sk.ca/about-the-assessment/>

⁴³ Agricultural data from Statistics Canada, retrieved Aug. 11, 2014 from:

<http://www.statcan.gc.ca/pub/11-402-x/2011000/chap/ag/tbl/tbl04-eng.htm>

replaced with other assessment instruments. There has been a province-wide survey of student attitudes about schools called “Tell Them From Me” (TTFM) introduced in these same two years without AfLs, but this is not related directly to academic outcomes and it is not intended for use by parents, schools, or individual teachers.

So using ministry information as my guide (rather than suppositions about what may be happening behind closed doors) Saskatchewan will test students in the early years (as mentioned above) and has in the years prior to my research also English in grades 4, 5, 7, and 10, and math in grades 5 and 8. The tests being piloted in selected school divisions from Pre-K through grade 3 were developed by a working group of representatives from four school divisions, the ministry, and interestingly, in part by outside agencies: (a) the Canadian Council on Learning; and (b) consultants from a firm called 'bv02.' The description from the ministry makes the assessments sound quite interesting, related very much to teacher reactivity and with clear feedback for educators and caregivers. One yet wonders how clear and strong these links will be to teachers. The assessments claim to be holistic, based on various types of data collection, and intended to “help create real and measurable change in the development of mathematical processes for children across Saskatchewan.”⁴⁴ They are still being piloted.

The only other academic assessments provincially mandated are grade 12 departmental exams for core subjects (English, biology, physics, chemistry, and three streams of math, but not calculus) when the local teacher is not accredited. Accreditation is a credential applied by the ministry which means, “Granting to a teacher the responsibility of determining the final mark or standing of the students in a specified grade 12 (level 30) subject or subjects.”⁴⁵ To avoid the need to give departmental exams, teachers can attend accreditation seminars which are offered every year and are intended to examine instruction and assessment strategies.

Since the newer LSAs are yet in the pilot stage, there is very little in the way of detailed information about the purposes to which the data will be put. They appear to be designed to provide timely feedback on student learning, to improve policy-level and classroom-level data-based decision making, and to promote individual interventions and strategies for improvement. These are five of the nine purposes determined by the researcher. The sixth refers to departmental exams as graduation requirements, and references to PISA and PCAP (the Pan-Canadian

⁴⁴ This reference to the Math assessment is from the Saskatchewan Ministry of Education, retrieved Aug. 11, 2014 from: https://www.edonline.sk.ca/bbcswebdav/pid-77653-dt-content-rid-821440_1/orgs/1401_biweeklybulletin_kleinkar/HMTAM%20-%202014-15%20Overview%20-%20May%202014.pdf

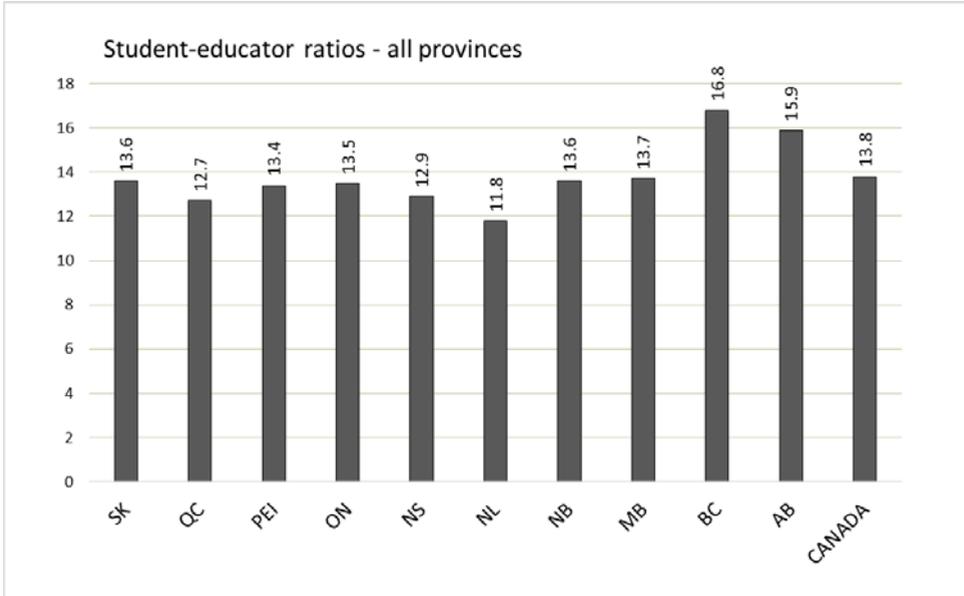
⁴⁵ Saskatchewan Ministry of Education, retrieved Aug. 11, 2014 from: <http://www.education.gov.sk.ca/accreditation>

Assessment Program) were made in the announcement of the Student Achievement Initiative ⁴⁶ which seems to be the starting point for all the changes in assessment policy that have occurred since that time (the cancellation of AfLs, the introduction of the TFM 'perceptual survey', and the piloting of new holistic assessments for grades 1-3).

⁴⁶ Saskatchewan Ministry of Education from May, 2012, and retrieved Aug. 11, 2014 from: <http://www.education.gov.sk.ca/student-achievement-annoucement-backgrounder>

1.6 Charts

Figure 1.7: A list of all provinces by SER. The national average is 13.8, and only 2 provinces are higher than that average figure.



Methodology

2.1 Introduction

Any national study of educational practices needs to be designed and carried out in a careful manner in order to reach the high standards expected of research in this field. The author has attempted to reach these standards. It is still the case that there remains plenty of room to disagree with the choices made, to dispute the conclusions drawn, but this will not be caused by any intentional lack of transparency intended to deceive the reader. All known limitations are acknowledged in this chapter or in Chapter 8 (pg. 306) on teacher background factors.

This chapter is laid out in the following way: (a) the theoretical framework is explained and related literature discussed; (b) educational reactivity is defined and referenced in the literature; (c) the author's unique reactivity model for education is provided; (d) the methodology of the study is explained and referenced in the literature; (e) the survey data collection methods are set out; (f) interview data collection is discussed in terms of its worth to the overall study; and (g) the specifics of the Canadian context of the study are examined.

2.2 Mixed methods

This study employs both qualitative and quantitative research methods (henceforth referred to as 'mixed methods') in order to seek explanations and justifications from actors across the educational hierarchy in several Canadian jurisdictions. Mixed methods seemed best suited to the researcher from the proposal stage to help make clear the complex process involved in educational assessment. Mixed methods were adopted to meet specific "theoretical, methodological and practical" advantages (as suggested by Brannen, 2007). Mixed methods are also, in this study, pragmatic and the approach tends to favour mixed methods work as well:

. . . pragmatic researchers are in a better position to use qualitative research to inform the quantitative portion of research studies, and vice versa. For example, the inclusion of quantitative data can help compensate for the fact that qualitative data typically cannot be generalized. Similarly, the inclusion of qualitative data can help explain relationships discovered by quantitative data. (Onwuegbuzie & Leech, 2005, p. 383)

I have used the quantitative data (surveys) to make comparisons across jurisdictions and the qualitative data (interviews) to illuminate the connections between these variables.

Using quantitative methods for survey data analysis, while providing numerical data for my research question, has its basis in some subjective choices – placing numerical values to concepts that defy easy quantification (in my case, mainly reactivity effects). Teacher responses to survey questions were given on Likert scales (ordinal measures) and have been translated into cardinal values (usually a range from -1 to 1) based on the judgment of the researcher (examples of this process are provided before regression tables in each following chapter). Onwuegbuzie and Leech (2005) see the quantification of 'abstractions' which are indirectly measured not as a flaw in quantitative methods, but as just another way that they are closer to qualitative than methodological purists would like to admit. There exist models that marry qualitative and quantitative methods, and Flick (2006) cites four distinct pathways which are mapped out by Miles and Huberman (1994). Three of these pathways suggest simultaneous qualitative and quantitative data collection, but the fourth pathway maps out a research design with a substantial time discrepancy in getting my survey and interview results (interview respondents identify themselves through first contact with surveys). I do not, as Miles and Huberman (1994) suggest, design a follow-up experimental study; however, it is certainly true that my 'complementary field study' is used to add depth to the data.

Beyond the fixed paths that can use qualitative and quantitative together, there are more general means identified in Flick (2006, p. 64) based on the work of Bryman (1992):

The logic of triangulation (1) means for him [Bryman] to check for examples of qualitative against quantitative results. Qualitative research can support quantitative research. . . both are combined in or to provide a more general picture of the issue under study (4). . . the problem of generality (7) can be solved for quantitative research by adding qualitative findings, whereas qualitative findings (8) may facilitate the interpretation of relationships between variables in quantitative data sets.

I would make the case that mixed methods means enriching the quantitative results while providing rigour to the qualitative data in my research design. It is worth noting, for the sake of caution, that Bryman (2007) years later re-visited integrated methodology work and acknowledged that many mixed methods

researchers did not truly integrate these methods in writing up their research. His caveats are well worth keeping in mind for mixed methods studies.⁴⁷

There are three potential positive outcomes from mixing methods in this way: (a) corroboration between data sets; (b) elaboration of thin datasets; and (c) complementarity of the datasets. A potential negative outcome is contradiction, and though that is an entirely possible outcome (Brannen, 2007), the richness of the study depends on collecting quality data. Researchers may make an error in thinking mixed methods will always lead to positive synergies.

Quantitative work in this field includes many studies that examine LSAs from the perspective of their results (Fuchs & Wössman, 2006; Koretz, 2002; Cartwright, Lalancette & Mussio, 2003). The focus, even when uncovering flaws in the testing models, is fundamentally on test scores and their value. My chosen theoretical model describes the educator's role in the LSA process and thus goes beyond any analysis of the test results (including value-added methods). It is problematic to use LSA results data to evaluate the assessments themselves because the value of the test results can best be considered in relation to other 'acknowledged truths' in the field of education, and they cannot be considered without appropriate context.⁴⁸

Another danger in using only quantitative methodology in this kind of study is one that is, in my experience, too much ignored economists writing on education (Hanushek & Rivkin, 2012; Lyons & Algozzine, 2006; Bishop, 1997; Bishop, 1998). Since so much research has been done on how standardized testing models are victim to fraud, misinterpretation and controversial achievement gains (Koretz & Jennings, 2010; Darling-Hammond, 2003; Crocco & Costigan, 2007), any quantitative analysis of achievement data is subject to substantial criticism related to the assessment tools used in the first place. In short, one cannot build a solid argument about the value of tests on questionable achievement scores.

Quantitative methods have been extensively used to try to show the value of LSAs, but clearly these methods alone are unable to provide much insight into how tests are perceived and best used by professional educators. With new leaks in

⁴⁷ Bryman (2007) identified nine factors which may deter truly mixing methods in final research reports: different audiences; having more faith in one methodology; study structure might favour one approach; different timelines for results; lack of specialist knowledge in both methodologies; belief in 'more interesting' data from one of the two; the difficulty in writing a coherent story with very different perspectives; publication biases; and finally the lack of a set of best practices shared by mixed methods researchers.

⁴⁸ James (1997, p.107) states that a pragmatic view of data means: ". . . they will be true, for pragmatism, in the sense of being good for so much. For how much more they are true, will depend entirely on their relations to the other truths that also have to be acknowledged." It is these 'other truths', or one of them at least, I wish to address.

the boat at every turn (teachers question the reliability⁴⁹ and see negative unintended consequences⁵⁰ from assessments), the use of LSA scores at face value would put in danger the arguments made here, and never get one to the topic of interest: instructional change. The effects of reactivity might be better determined by side-stepping LSA scores and collecting data directly from teachers using a combination of surveys and interviews. Since the questions for interviews were based upon survey responses, the method was inductive with the use of the data gathered from survey respondents to determine working hypotheses to be addressed in interviews (Blaikie, 2006; Thomas, 2006).⁵¹ By aligning these two data sources (triangulation: from Jink, 1979) and using the strength of personal experiences to give colour to the regressions from survey data (mixed methods: from Onwuegbuzie & Leech 2005), these disparate aspects of the study combined to provide informative correlations as well as interesting anecdotal insights.

2.3 Surveys

Since many themes related to large-scale assessment have been extensively covered in the current literature, these themes informed the design of the survey. Data collection was employed to confirm if the experiences of teachers in other jurisdictions hold true in Canadian contexts. Follow up interview questions were written to target the same themes as the phase-one data collection, and potential respondents indicated their willingness to proceed with interviews within the survey itself.

The survey was written using Survey Monkey which tabulated directly into a spreadsheet (reducing potential errors; Nardi, 2002) and which was password protected by the author. Surveys were sent to teachers in all ten Canadian provinces, and to educators working in schools at all grade levels K through 12. The survey design was approved for circulation by the Maastricht University GPAC² Supervisory Committee in Oct. 2013. Many of the questions had

⁴⁹ "The testing infrastructure on which so many school reform efforts rest, and in which so much confidence has been vested, is unreliable - at best." (Finn and Petrelli in Ravitsch, p. 107)

⁵⁰ "High-stakes tests are often associated with other unintended consequences. They include retention of students in grade before tests; suspension, expulsion, and reclassification of students before tests; "teaching to the test;" the narrowing of the curriculum; the loss of teachers from the profession; and cheating." (Amrein & Berliner, 2002, p. 35)

⁵¹ Further clarified by Thomas: "The primary purpose of the inductive approach is to allow research findings to emerge from the frequent, dominant, or significant themes inherent in raw data, without the restraints imposed by structured methodologies." (Thomas, 2006, p.238)

been field-tested in well-regarded and related research studies, and were therefore considered to be valid measures of reactivity effects. Questions about reactivity practices, the real core of this study, were adapted from both Skwarchuk (2004), and Hamilton and Berends (2006). The format was guided by the example survey from Kemp and Freisen (2009). Questions on teacher attitudes were adapted from Brown (2004). Questions on appropriate uses of the data were adapted from Wayman, Cho, Jimerson and Spikes (2012). Questions related to supports and professional development were influenced by Boyle, Lamprianou and Boyle (2005).

Following the guidelines of the Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada (2010) commonly called the Tri-Council Policy, this survey qualifies as "program evaluation" which falls outside the ethics guidelines of the more involved "human research."⁵² Stock and Watson (2007, p. 469) also make this same distinction clear. Thus no Ethics Board approval was granted or sought.

Email-based surveys are considered an effective and efficient means of collecting basic information from large groups of respondents (Selwyn & Robson, 1998; Flick, 2006). They also offer the advantages of low-cost distribution, response rates comparable to paper-based surveys, and wide coverage of the sample (Couper, 2000; Evans & Mathur, 2006).⁵³ There is also the potential to reduce coverage and sampling errors since all teachers in the public school system in Canada have daily and regular access online and email accounts.

Flick (2006) speaks of the trade-off between wide coverage (like these surveys) and in-depth data. My goal is to get data that examine reactivity effects in general and across jurisdictions with the surveys, while deferring to interviews for the in-depth information gathering.⁵⁴ The mixed methods approach hopefully

⁵² "Article 2.5: Quality assurance and quality improvement studies, program evaluation activities, and performance reviews, or testing within normal educational requirements when used exclusively for assessment, management or improvement purposes, do not constitute research for the purposes of this Policy, and do not fall within the scope of REB [research ethics board] review. Application: Article 2.5 refers to assessments of the performance of an organization or its employees or students, within the mandate of the organization, or according to the terms and conditions of employment or training." (Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, 2010, p.20)

⁵³ Although as web surveys have become more common, response rates have apparently decreased (Sax, Gilmartin & Bryant, 2003).

⁵⁴ As noted by Boyle, Lamprianou & Boyle (2005), surveys report reflections without further analysis: "Another limitation of the study is the self-reported nature of the information that is provided by the respondents. The findings of this study are valid to the extent that the self-reported information is accurate. The accuracy of self-reported personal or other

avoids this trade-off with a high 'n' for survey data, and high quality individual responses from interviews.

Onwuegbuzie and Leech (2005, p.383) equate the combined use of qualitative and quantitative data sets as a means of achieving two important and complementary goals: "empirical precision with descriptive precision." Surveys have the advantage of being suited to examining attitudes about LSAs, since these are not directly observable, and allow teachers to respond honestly with the advantage of anonymity (Nardi, 2002). This method of research has proven to be effective in eliciting specific information about data use in a study by Anderson, Leithwood and Strauss (2010) where surveys were followed up by interviews of chosen subjects (based on *relevance* of their surveys responses).

2.4 Survey sampling

This was a cross-sectional study design – the surveys were all dispersed in the same school year across Canada, and only once to any one school or school division. The sampling unit was individual teachers, the target population was Canadian public school teachers, and the clusters for the probability sampling were school divisions. Using cluster methodology, this target population was already geographically divided into non-overlapping groups (school divisions). School divisions (also called boards or districts) were randomly selected as clusters in each province. Each school division (cluster) was comprised of a variety of strata (experience, grade level taught, qualifications, etc.) and schools of all sizes and grade ranges were included. All teachers from participating schools and divisions were asked to take part. This type of sampling gave access to a heterogeneous group of respondents as participant school divisions were sought out until a minimum number of responses was gathered. Cluster sampling, being specifically targeted, in some cases can decrease the coverage errors of a large census frame (Fricker, 2006).⁵⁵ This sampling provided both a meaningful and a manageable amount of data from teachers in different administrative clusters which all have potentially different LSA policies (set in schools or divisions).

information has been extensively researched in the past and is still researched in the context of educational and social research." This sentiment is echoed in Nardi (2002) who adds that ". . . behavior is a much better indicator of what people feel or think about a subject. . ." than questions *about* their beliefs.

⁵⁵ Fricker (2006, p. 35) ". . . researchers who only focus on reducing sampling error by trying to collect as large a sample as possible miss the point that it is equally important to reduce coverage, measurement, and nonresponse error in order to be able to accurately generalize from the sample data."

It is worth noting that the selection process was in name random, but was built on the practical reality that many school divisions chose not to take part. This made the cluster choices seem more like a voluntary sample than a random one. While original applications were made randomly (attempting simply to match provincial urban and rural proportions since school divisions, as noted above, include heterogeneous populations of teachers), inclusion of any given division was far from certain and in many cases further applications were necessary. Teachers in divisions that did not distribute should not, therefore, be classified as non-respondents since they never had an opportunity to take part. They were simply non-participants. This same dynamic was true of school administrators; they were in all cases allowed the discretion to opt out of distributing the survey within their schools, so teachers in these schools were likewise non-participants. Thus both division- and school-level administration served as gatekeepers allowing or denying survey distribution based on their preferences and value judgements.

The result of the survey distribution was a set of responses from a wide cross-section of educators based on those pre-determined criteria which proved valuable for later analysis without an immense random sampling being necessary or setting prescriptive conditions about who gets emailed the survey. Comparable methods were employed by Banicky and Noble (2001) for a single-state case study of teachers' opinions of large-scale assessment, and also in Taras (2004) for identifying teachers' opinions on state testing programs in Florida. In each case, the state sample (or cluster) was seen to be largely representative of the nationwide population of educators, with variances based on state policies and conditions.

There are several important strata in my target population: (a) the teaching population includes teachers from urban locations as compared to those from rural areas; (b) teachers in large schools or small schools; (c) teachers at different grade levels; (d) teachers of different ages and sexes; (e) teachers who teach different subjects; (f) teachers with more or fewer years of experience; (g) teachers with different qualifications; (h) teachers with larger or smaller classes; and (i) teachers from all ten provinces. The provincial samples are smaller, but the larger 'n' of the combined data means that all of these strata are represented in the national data. National data are only available for two of these strata (these being sex and age of teachers), but a comparison of the collected sample to Statistics Canada data shows that they are very much the same in the only nationally comparable data (see **Figures 7.2 and 7.3** on the congruence of sample and target population age and sex data as well as the comparison methods used).

There was no issue of coverage errors between the target population (in-service public school educators) and frame (school district staff) since all public school teachers in Canada must belong to a school division. Excluded divisions (namely Catholic divisions and private schools) can, in some cases, have unique curricula, non-standard procedures and different governance policies than public

schools. This would have made their inclusion in the provincial samples problematic. They do not, however, represent a significant proportion of the Canadian student or educator populations.⁵⁶

This was not the case regarding the French-speaking populations in the provinces of Québec (80.0%) and New Brunswick (28.4%). These two provincial French populations make up fully 92% of Canada's Francophones.⁵⁷ In all provinces French-language school divisions were excluded, but this exclusion clearly represents a larger proportion of students in both Québec and New Brunswick. This choice can be rationalized in that these provinces have separate (but both parallel and equal) English school boards which were the ones approached for this study. Assuming, as they profess, that the English- and French-language boards follow the same assessment policies, drawing a random sample from only English-language schools should not introduce any bias and would circumvent the not inconsiderable difficulties of translation (for surveys and interviews).

2.5 Other potential errors

Measurement errors were addressed by running a thorough and diverse pilot of the survey to eliminate possible comprehension problems, poor wording or design, technical flaws, etc. (Couper, 2000). Non-response errors were minimized by keeping the survey simple and short (Ray & Tabor cited in Evans & Mathur, 2006). Other design aspects which might improve the quality of the data (as in Fricker & Schonlau, 2002; Nardi, 2002) were also incorporated such as avoiding sensitive topic items and reducing the number of open-response items. Nardi (2002) also notes that while closed response questions allow for less variation in responses, they make the survey faster and easier to complete than typing open-response answers. Surveys that are longer or that are difficult to complete tend to have lower response rates (Brown, 2003 cited in Evans & Mathur, 2006). Closed response questions also make coding and categorizing responses more efficient for the researcher and considering that follow up interviews were always an integral part of the research design, it was thought to be a good trade-off of survey length for ease of completion.

⁵⁶ "The legislation and practices concerning the establishment of separate educational systems and private educational institutions vary from jurisdiction to jurisdiction. . . Public and separate school systems that are publicly funded serve about 93 per cent of all students in Canada." (from the Council of Ministers of Education, Canada website: <http://cmec.ca/299/Education-in-Canada-An-Overview/index.html#03>, retrieved Aug.2014). Statistics Canada most recent figures sets the number at a comparable 93.2%. (<http://www.statcan.gc.ca/daily-quotidien/010704/dq010704b-eng.htm>)

⁵⁷ From Statistics Canada: 2011 Census Population (by language spoken), see **Figure 1.4**.

Sax, Gilmartin and Bryant (2003) explain that one particular error, non-response bias, is particularly difficult to account for since even a low response rate does not necessarily indicate that non-response bias is present. Only if non-responders have different opinions than responders on the survey topic is this bias an issue.⁵⁸ Assumptions about non-responders are likely not wise to make, and answering to the issue a better solution.⁵⁹ Stang and Jöckel (2003) note that twisting the arm of someone who really is not interested in responding to a survey, and thus a respondent who is less likely to take the time to do so with any accuracy, may lead to skewed results that undercut efforts to get accurate data from the sample.

The most effective method (eventually) found by this researcher to deal with non-response was to improve response rates, and this was accomplished through personal contact with school administrators. In my first survey distribution following the pilot, the researcher spoke to a division superintendent who agreed to forward the email link to all school principals in her division. There was no way for the researcher to verify that the administrators had, in fact, forwarded the survey link to their staff. As a result the response rates were very low (4.37%). After this experience, surveys were never sent to schools without contact being made with a school administrator. This way the researcher was able to gauge the level of interest shown by the principal, to see if the principal was willing to send three emails to all teachers (numbers of whom were measured with a common metric – full-time equivalences or FTEs which counts only professional staff and is almost universally used) over five school days, and thus I was able to avoid those schools which had that little apparent interest or time and also to ensure that teachers had ample opportunity to take part. Teachers were not contacted directly.

Following the adoption of this procedure, response rates were never below 10% (the working 'margin of acceptability') for any division, were regularly over 20%, and often closer to 30%. The national figure, which includes all school and divisional aggregated data, has a response rate of 19.4%, with more than one in three (42.3%) of the respondents indicating that they give LSAs in their classrooms. Higher response rates and the congruence of the sample population to the national

⁵⁸ Olsen (2006) relates: "However, recent research (Curtin, Presser, and Singer 2000; Keeter et al. 2000; Merkle and Edelman 2002) has called the traditional view into question by showing no strong relationship between nonresponse rates and nonresponse bias (Groves 2006)."

⁵⁹ "While the assumption that nonresponders either all have or all do not have the characteristic in question stretches belief, the usual assumption that responder characteristics do not differ from nonresponder characteristics is also implausible. In most surveys, there will be differences between responders and nonresponders and even modest differences will lead to large biases unless nonresponse is very low." (Colombo, 2000, p.85)

population on the only two nationally available statistics help answer the question of non-response in these data (Fricker, 2006; Saxon, Garratt, Gilroy & Cairns, 2003; Olsen, 2006). **Table 2.1** shows the figures on respondents, participating divisions and schools, as well as responses rates province to province and nationally. Surveys were completed between June 2013 and June 2014.

Table 2.1: Responses to nation-wide teacher survey. For analysis, partially-completed surveys were removed from these respondent groups.

Prov.	Number of participating divisions	Number of participating schools	Participating schools' total FTEs	Total responses	Responses from teachers giving LSAs	Total response rate (%)	Respondents giving LSAs (%)
AB	4	18	561.4	118	48	21.02%	40.68%
BC	4	32	808.9	75	43	9.27%	57.33%
MB	7	29	669.2	130	40	19.43%	30.77%
NB	2	18	649.2	151	59	23.26%	39.07%
NL	1	13	313.9	73	30	23.26%	41.10%
NS	2	23	335.1	92	61	27.45%	66.30%
ON	7	25	630.9	108	52	17.12%	48.15%
PEI	1	28	568.4	92	35	16.19%	38.04%
QC	3	13	302.6	62	30	20.49%	48.39%
SK	4	40	683.7	170	55	24.86%	32.35%
CANADA	27	181	5523.1	1071	453	19.39%	42.30%

2.6 Survey data analysis

The survey data were analyzed using Stata software. Important steps in the process (guided in large part by Mitchell, 2010) included cleaning the data (for example, some 'other' responses were really explanations of choices), labelling datasets, creating variables (reactivity effects scores were created from the distinct responses), and combining datasets (from different divisions, and provinces).

These data were analyzed using four main strategies which grew more involved at each level. First, count data were done to describe the data set and the different variables included within it. Rating scales and survey answer options were included in this section to make clear what choices were presented to respondents. These most basic analyses are found in the chapter-ending sections.

Next, distribution analysis was done where there were sufficient numbers of respondents and answer choices, and when the data follow the general pattern of a normal distribution (i.e. where this type of analysis has some merit).⁶⁰ All basic

⁶⁰ Some distributions (the 'years of experience' background variable, for example) do not meet this criterion because each group within the distribution may have approximately the

statistics on the distributions are provided (the mean, standard error, variance, range of responses, skewness and kurtosis, and the number of observations from which the distribution was derived). The distribution value (“D” is the index of variance, also called the VMR - variance to mean ratio) is also used to determine under- and over-dispersed curves. The ratio of the population variance (σ^2) to the mean (μ) (thus, $D = \sigma^2 / \mu$) determines this value. D values higher than 1 indicate over-dispersed curves, values lower than 1 indicate under-dispersed curves, and values at or close to 1 indicate more normally dispersed curves. While under- and over-dispersed curves can be normally distributed, this value gives some insight into which part of the scale the bulk of the responses were given.

Correlation matrices were next created to examine the relationships between independent variables prior to completing the regressions. This was one means of determining if there were covariant variables in the study, and also to see which apparent close relationships between independent variables stand up after introducing the dependent variables into the mix.

In the final analyses the dependent variables are explained with a set of independent variables in multiple regression models using ordinary least squares estimation. The dependent variable (Y) in my study is teachers' use of LSA data. This dependent variable was determined based on a series of questions asking what uses teachers have made of LSA data in their classrooms. Independent variables (X) will include four main lines of inquiry operationalized for the survey into numerical values: (a) test data and design; (b) supports provided for teachers; (c) incentives for teachers to use these data; and (d) teachers' attitudes regarding the tests. The multiple regression equation, which follows the assumptions of the OLS for multiple regressions, is as follows (from Stock & Watson, 2007):

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

- Y_i 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 X s equal 0
- β_1 is the regression coefficient of X_1 ; β_2 is the regression coefficient of X_2, \dots, β_k
- e_i is the residual of the regression, analyzed following each regression for its normal distribution, heterogeneity, and other potential abnormalities

same proportion of respondents. Some questions have only two or three possible answers (the ‘school setting’ variable is an example where urban/suburban or rural/remote are the only choices), so distribution analyses here are not suitable.

The analyses (and tables) were done in two steps: independent variables were added and computed; next, provincial dummies were added. Nine dummy indicators added to partially account for variations between the ten provincial jurisdictions. In some cases provincial samples were too small for independent analyses, thus introducing these dummies allowed for some of the differences in provincial testing systems to be explored. Adding provincial dummies, each province also being one stratum in the national sample, was intended to account for differences in selection probabilities between provinces as an alternative to weighting. The sample is comprised of many strata some measured in different ways in different provinces (such as qualifications being measured by ten unique certification systems, urban areas are determined based on different population levels, etc.), it would have been difficult to gauge appropriate weights. It should be noted that the one provincial dummy left out of the regressions, the control if you will, changed depending on the reactivity type examined. For each type of reactivity, positive, negative, and total, the province with the reactivity score closest to the national average for these effects was left as the control. In this way variations in both directions (positive coefficients and negative) might be noted and examined. So PEI is the control province for total reactivity, BC for negative reactivity, and MB for positive reactivity (these scores are in **Table 3.3**).

Four additional checks were done on all the regression outputs in order to confirm the robustness of the results. Robustness checking methods were used for all the regressions done (fifteen tables in total) and the results of only a representative sample are presented in the interest of space. The graphs are: (a) observed values plotted against the predicted reactivity values from the regression (observed values v. \hat{y}); (b) the estimator of the expected value of x compared to the predicted value of the regression model (\hat{e} v. \hat{y}); (c) a histogram of the residuals plotted against the normal distribution; and (d) a QQ plot that compares the quantiles of a regression variable or a residual with the quantiles of a normal distribution. In this thesis the *qnorm* function in Stata statistical software is used, which is a specific type of QQ plot which focuses more on the distribution of the tails. These residual analyses are discussed in the body of the chapters, and the charts themselves are available in the chapter-ending sections.

Each of these methods was used to check different aspects of the regression model. (a) The 'observed v. \hat{y} ' graph was done to determine if there was a linear relationship present between regressed independent variables and the dependent variable. In strong and significant results a positive linear relationship should be apparent and, ideally, with no apparent clustering. (b) The ' \hat{e} v. \hat{y} ' comparison was done to determine heteroscedasticity, and if there was noticeable clustering, non-linear relationships, or serious outliers that might unduly influence the regression results. (c) The histogram allowed a check on the distribution of the residuals. It is not an expectation that they be normally distributed, but the closer

they are to the normal distribution curve, the more likely it is that they meet the independent and identically distributed (i.i.d.) condition of OLS regressions. (d) Finally, QQ plots of the quantiles of a variable against the quantile of the normal distribution were used to evaluate the distribution of residuals. As mentioned above, the Stata '*qnorm*' test is a variation on the QQ plot that examines the tails in more detail. This is beneficial in this case in light of the fact that tails in these distributions can harbour outliers or be truncated (especially in the cases where discrete variables are used, like this study).

A final analysis done was the use of Cronbach's alpha to gauge the internal consistency of both positive and negative reactivity questions (Tavakol & Dennick, 2011). It is important in the case of aggregated scales (positive and negative reactivity scores are aggregated from several survey responses) to determine how much the survey items appear to measure the underlying construct (Santos, 1999). Cronbach's alpha is this sort of index of reliability, but even this index can be over-estimated in value by some researchers (Sijtsma, 2008).

It should also be noted that research into multilevel analysis methods (also called hierarchical linear modeling) which was done after the distribution of the survey proved informative, and the survey design would have been different had this research been done earlier. Multilevel methods are intended to address the likely violation of the i.i.d. assumption found in most basic statistical methods (Hox, 2006; Garson, 2013). The basic unit of my study, teachers, are nested within their schools, and multilevel methods recognize that the culture of the school is likely to make teachers from the same school respond similarly, and thus violate the i.i.d. assumption. In designing the survey, the researcher was more focused on guaranteeing anonymity to respondents so that they might answer honestly, and thus no school identifiers were used. In hindsight, using a school identifier would have made the analysis via multilevel methods possible, but also might have prompted rejections from both privacy- and data-sensitive school divisions. Without the identifier, this form of analysis comes down to OLS.

Comparing provinces with different assessment programs will introduce the possibility of external validity concerns – assessment policies may differ enough across Canadian provincial jurisdictions to make comparison and externally valid generalizations impossible (Stock & Watson, 2007). Manitoba, specifically, has a very different-looking provincial assessment that is given in elementary and middle years grades (this is noted in Chapter 1). Aside from this unique (and interesting) evaluation method, it can be seen in **Figure 1.1** that the subjects and grades tested are quite different from province to province.

There is no standard testing model for provincial large-scale tests, therefore, a comparison across provinces is complicated. Yet I would argue this makes a national study all the more relevant in examining the methods and purposes of assessment across all ten jurisdictions and how these align with

teacher practices. Surveys included numerous respondents from all ten provinces, and the purposive sampling methods for interviews meant that at least one interview respondent came from each of the provinces. The intention of purposive sampling here was to ensure that all forms of reactivity (positive, negative, and neutral) were addressed.

2.7 Interviews

Follow up interview questions were written to target the same themes as the survey from a sample of respondents who had indicated their willingness to proceed with interviews. These interviews were done as a means of triangulating quantitative and qualitative data sets. Jink (1979, p. 602) identifies triangulation as "a vehicle for cross validation when two or more distinct methods are found to be congruent and yield comparable data." Flick (2006) advocates the purposive selection of respondents (purposefully and based on relevance to the topic) as a means of triangulating follow up interviews with the results of preliminary data gathering.⁶¹ This is triangulation in the sense that qualitative and quantitative methodologies are combined, and not in the other sense of the word, which is quite different, where the researcher combines several qualitative methods. Purposive sampling was used to gather a relatively balanced number of responses from all three types of reactivity effects (interviews included teachers who reported positive, negative and neutral reactivity effects). Starting from a list of self-identified willing respondents, the purposive selection process ensured: (a) having willing subjects; (b) having subjects knowledgeable about the topic; (c) getting detailed responses from all three kinds of reactivity practitioners; and (d) having a selection of subjects from all provinces in Canada. Numbers were determined in order to set a logical scope – not too large to be impractical, nor too small to provide meaningful results (Gerring, 2012).

Interviews were conducted via telephone since distances were in almost all cases prohibitive. Interviews were semi-structured and were conducted using elements of two specific methods: (a) the expert interview which sets out to get specific and targeted information (Flick, 2006); and (b) the ethnographic interview which sets the data in a context of the culture (Spradley, 1979). In both these methods, it was beneficial to share common pedagogical background and

⁶¹ "The same people are interviewed and fill in a questionnaire. Their answers in both are compared with each other, put together, and referred to each other in the analysis. . . The same people are included in both parts of the study, but in a second step, it has to be decided which participants of the survey study are selected for the interviews." (Flick, 2006, p.37)

knowledge, but also to require explicit explanation of provincial variations on the LSA theme.

Coding of the interviews was done using the established independent variables and themes of the survey. Since the intent was triangulation, interviews were clearly meant to plough the same fields but to greater depths. A copy of the coding key can be found in **Annex 4** and it was built on the premise (from Flick, 2006, p.67) that researchers should select the “coding procedures [which] seem to be the most appropriate.” Since the independent variables were set for the survey, it seemed logical to code using these same variables. Thus the process was more like axial coding than open coding - categories relevant to the research were pre-determined (Flick, 2006). Blaikie (2000, p. 62) also speaks positively of an efficient coding process “if coding frames are established before the data collected, such as in a questionnaire.”

Interviews were completed by October 2014, so the time-lag between completing the survey and follow up questions was not excessive. Assessment policy is something of a moving target in Canada, with many provinces regularly changing what they test, when they test, and how they test it. This was another reason it was critical to collect all survey data within a single school year and to follow up in a timely way with interviews.

Figure 2.2: Review of the methodology literature

Topic	Author(s)	Summary statement
Mixed methods	Blaikie, 2006	Comprehensive text on qualitative methods.
	Brannen, 2007	Notes pressures to combine methods are getting stronger from the public and researchers themselves, and gives a specific example of study suited to complementary uses.
	Bryman, 2007	Interviews with mixed methods practitioners found that many claim to use mixed methods but do not integrate findings from qualitative and quantitative or only report on one set.
	Denzin & Lincoln, 2000	A history of and resistance to accept qualitative model in which interpretive paradigms are explained.
	Flick, 2002	An overview of current research on triangulation.

	Flick, 2006	Text covering all aspects of qualitative methodology: interviews, qualitative v. quantitative methods, coding, etc.
	Jink, 1979	Cites quite old examples of triangulation and mentions it can be done poorly, but has great potential when done well. An example from his own work is given which mentions in particular trying to explain a lack of convergence.
	Onwuegbuzie & Leech, 2005	Paper argues the divide between qualitative and quantitative is a false one built on faulty premises and that they are better seen as a continuum or exploratory and confirmatory studies.
Surveys	Anderson, Leithwood & Strauss, 2010	Examines organizational factors in data use across jurisdictions. This study has a similar design model to the author's but a different theoretical model.
	Couper, 2000	Web-based and email surveys are examined looking at response rates, sampling error, etc.
	Evans & Mathur, 2006	Exhaustive chart of the pros and cons of email surveys including some tips to overcome common issues.
	Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, 2010	The research ethics guidelines adhered to by Canadian universities, research organizations and public institutions
	Nardi, 2002	A basic text on surveys including some definitions and clarification of some models.
	Sapsford, 2007	An introductory text on surveys with lots of information about the statistical use of operational survey data.
	Trouteaud, 2004	Studies how to get response numbers up on email surveys. The conclusions are that follow

		up, pleading, and under-estimate the time it takes are effective.
	Vincente & Reis, 2010	Meta-analysis of web designed surveys includes information about on question design, structure, progress indicators, button types, amongst other things.
Sampling	Banicky & Noble, 2001	Delaware educational case studies which relate to reactivity, but cited for example of similar sampling model.
	Cochran, 1977	Technical text on sampling methods which is informative on clusters, basic analysis (distributions), and also nonresponse.
	Fricker, 2006	Strong examination of sampling: using a large sample does not correct for coverage, measurement, or nonresponse errors.
	Taras, 2004	Looks at a teacher's college HSA assessment and resulting changes to practices. This paper is cited as an example of similar sampling model.
Other potential errors	Fricker & Schonlau, 2002	Evaluates assumptions that online surveys are faster, cheaper, and better, concluding that they are not really (except cheap).
	Hudson, Seah, Hite & Haab, 2004	Environmental survey via mail, mail-internet, and internet finding nonresponse bias was not related to delivery method (same across all media).
	Sax, Gilmartin & Bryant, 2003	Looks at nonresponse issues with surveys and finds no strong predictor in study done with college students except gender (non-response a non-issue).
	Stang & Jöckel, 2003	Epidemiological paper on response rates, cited since successive recruitment waves become less reliable respondents (i.e. too much pleading makes for poorer data).
Data analysis	Duyar, Gumas & Bellibas, 2013	Turkish study using multi-level analysis (MLA) to examine level 1 teacher variables and level 2 school variables on the topic of teacher efficacy and school leadership. Interesting for its context and use of MLA, equations and regressions shown and layout is simple, clean:

		an example of MLA used in educational research.
	Ferron, Hess, Hogerty, Dedrick, Cromley, Lang & Niles, 2004	Technical piece on the common uses of multilevel analysis showing that studies using MLA do not have high standards for: deciding on predictor; covariance structures; cross-validation and sensitivity analysis; centering; dealing with residuals; estimation; dealing with outliers; missing data; identifying software or regression models; significance tests, etc. It ends with recommendation list for future research.
	Garson, 2013	Technical paper (especially when discussing software use) but key terms are defined and provides good examples used to discuss multilevel analysis.
	Hox, 2006	Covers why multilevel analysis is used, how it is used. To illustrate this, the author gives different examples and equations, shows regression tables, and integrates topics such as centering, examining residuals, etc.
	Mitchell, 2010	Guide to using Stata for managing data from importing files to outputs. This is a very helpful guide to practical use.
	Steenbergen & Jones, 2002	Strong arguments for the use of multilevel analysis (MLA) in political science and discusses the mathematics involved some. The justifications for use are strong, and superiority over non-MLA models is examined. It ends with a practical three-level case, modeling, equations, etc.
	Stock & Watson, 2007	Comprehensive text on econometrics with lots of material here on data analysis: regression models, threats to internal/external validity, nonlinear regressions, how to read/create regression tables, assumptions and limits of models used, etc.
Interviews	Gerring, 2012	A text on social sciences methodology which looks at aspects such as coding and sampling using a thematic approach (which makes it

		difficult to pick apart by topic).
	Hay-Gibson, 2009	Examines the use of Skype for interviews and lists the advantages and drawbacks.
	Spradley, 1979	Book examines the various methods of interview, examples from author's studies, and a focus overall on the cultural differences within groups in society.

2.8 Canadian context for interviews and surveys

The target respondents for surveys and interviews are seen in **Figure 2.3**. It should be said that the respondents with the least power in the educational hierarchy (aside from non-professional staff) are also the most important to this study: the teachers. Many of the officials listed in **Figure 2.3** have the authority to make and enforce policy choices, but in this study these provide only context. Only the teachers themselves can talk of how data are used in their classrooms, how they prepare their students for LSAs, and what pressures they feel are applied to them based on the results. Since this is what the study is all about – the reactivity of teachers to external evaluation – this data was central to the analysis and interpretation of literature, and interviews with other officials. Speaking to these same teachers' administrators, their superintendents, and their directors would allow for the data to be triangulated in the interest of validity (Boyle, Lamprianou & Boyle, 2005; Gerring, 2012).

Figure 2.3: Respondents to surveys and interviews

Respondent group	Epistemological interest	Tool	Numbers
Division level - directors - superintendents	- PD provision - data analysis and reporting - resource allocation - program improvement	Interviews	10 directors or superintendents
School level - administrators - teachers	- improved instruction - positive/negative reactivity effects - organizational strength - skills for working with data	Interviews Surveys	Interviews: 20 Teachers: 15 administrators Surveys: 300-400 teachers

At the school level, interviews were requested with several administrators from each province to establish what criteria might be set school-wide for LSA data use. They have direct contact with tests and policies and their expectations would likely be reflected in teacher reactivity. The best means of triangulation in this case was to choose administrators from those who work alongside teachers who had been interviewed.

Next up the hierarchy are divisional directors and superintendents. These officials have more input about how LSA tests are rolled out, how they are supported, and what expectations are established for the data at this level. Interviews were sought from at least one division-level employee from each province. These potential candidates were selected based on their division's participation in the research project. **Figure 2.3** indicates proposed numbers of respondents that were to be interviewed, and assumes that data 'saturation points' will vary from province to province.⁶²

2.9 Conceptual framework

A visual representation follows in **Figure 2.4**, it being the conceptual framework used in this dissertation and partially illuminated in this chapter. The three jurisdictional levels (provincial ministries, school divisions, and schools) all have policy goals and input that affect the practices of the classroom teacher. This is true of LSA policy amongst other initiatives. This policy input is sifted through the pedagogical filter of a professional code (in this case the STF Code of Professional Competence) and results in some form of reactive effect: positive, negative, or neutral. These reactivity effects may have an impact on LSA scores, and the data from these tests are then fed into the system again where policy decisions are made based on the results.

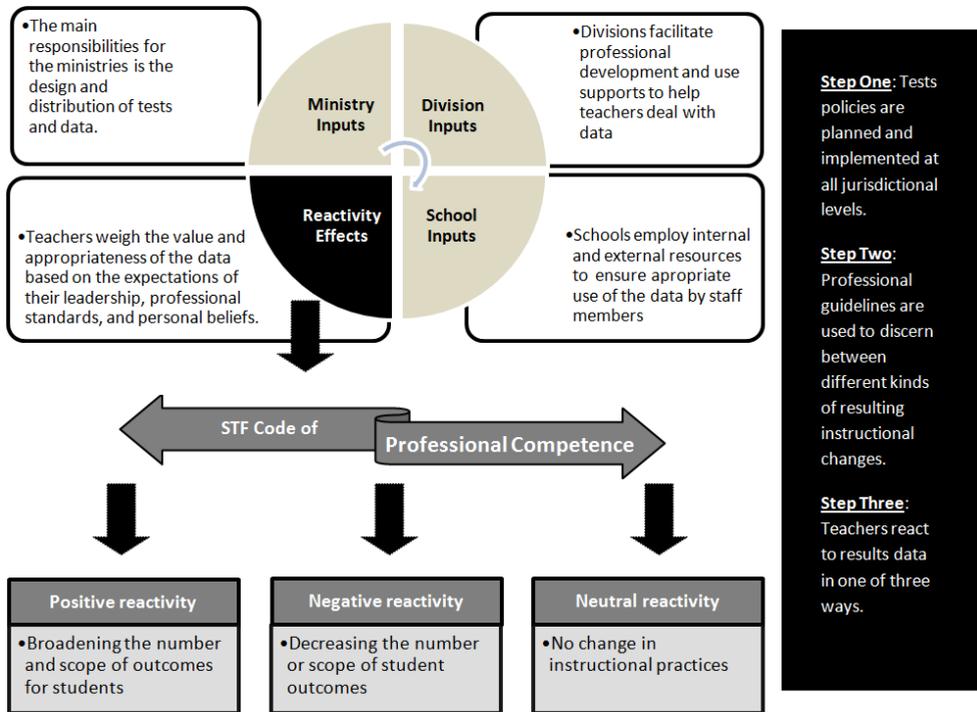
In general, all jurisdictions are seeking similar goals: (a) adherence to curriculum; (b) improved instruction and programming; (c) data-informed decision making; and (d) system accountability. Each level also has a mandate that dictates which policies they can create to help reach the system-wide goals. So while provincial tests are designed and built by the provincial ministry, strategies to prepare students for the tests and to align instruction to curriculum goals might well be division- or school-level policy choices.

In the end, there is a choice made by a teacher about how they will use the data to improve their teaching, and that is the key determinant of an assessment

⁶² Glaser & Strauss (1967, p.61) define theoretical saturation in Flick (2006): "Saturation means that no additional data are being found whereby the sociologist can develop properties of the category."

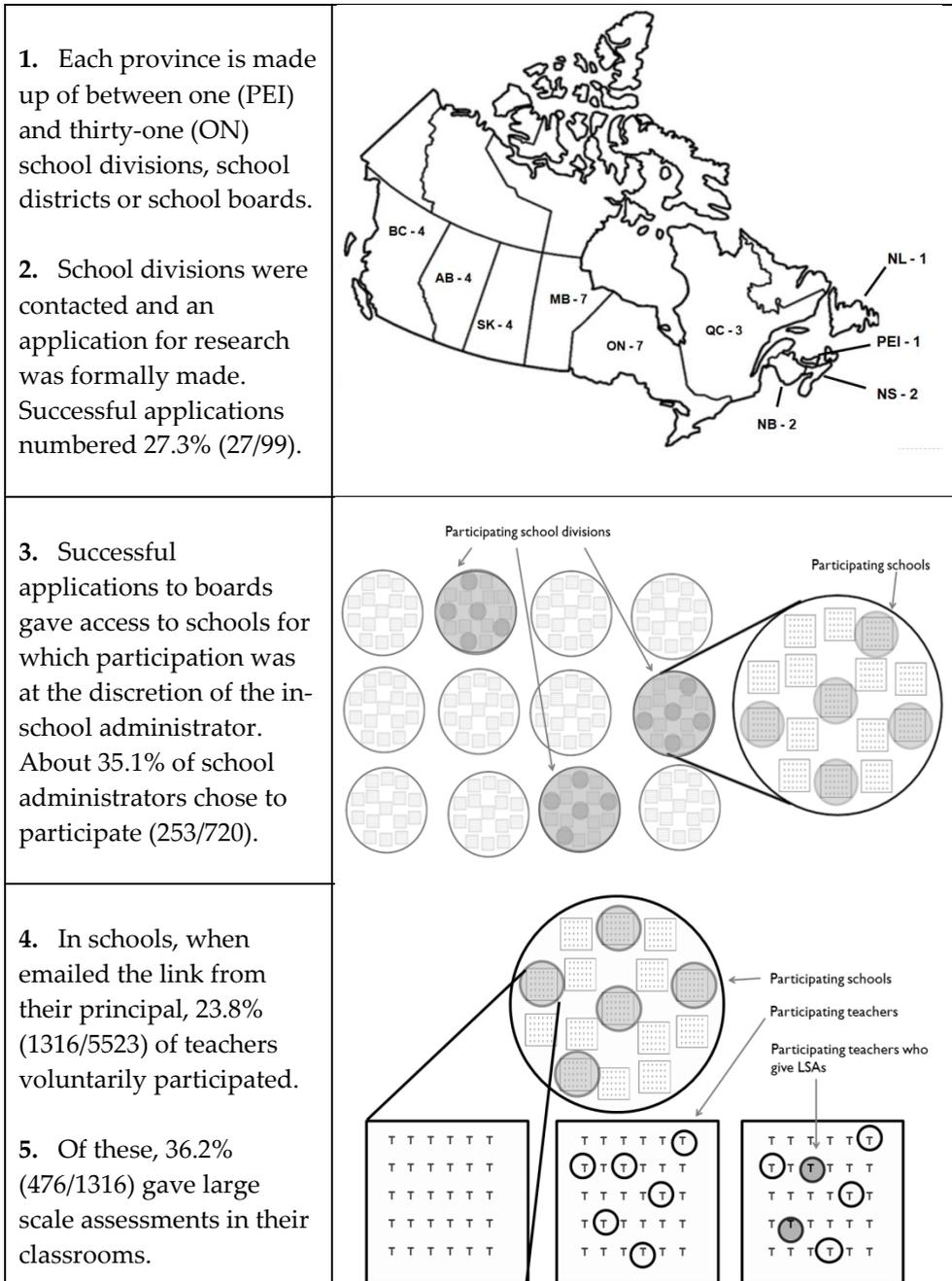
system is meeting the goals it has set. These individual choices, made in thousands of classrooms every day, are the heart of this study.

Figure 2.4: Conceptual framework for this study



2.10 Charts

Figure 2.5: Sampling method



Reactivity in Canadian education

3.1 Introduction

This chapter sets out the theoretical framework of the entire study: the concept of reactivity. Reactivity is the way people react when being assessed by an external method or source. An evaluation of what we are doing, when aware of the fact that we are being evaluated, changes our actions. This is true especially if it is known which aspects of our behaviour are being examined. Provincial large-scale assessment (LSA) is a form of program evaluation intended in large part to check the effectiveness of teaching, learning, and reporting these findings to the government body with oversight powers (usually a ministry of education). It stands to reason that reactivity in education is how teachers react to external evaluation of their work performance (and by extension students' work) by provincial ministries.

What follows is laid out in the following way: (a) the theoretical framework is further explained and related literature is discussed; (b) educational reactivity is defined and referenced in the literature; (c) the author's unique reactivity model for education is provided; and (d) results from national surveys related to reactivity (the dependent variable of my study) will be examined.

3.2 Theoretical framework

This study is based on the theoretical concept of reactivity. In social studies research, people react to being observed or evaluated, and these reactions, whether consciously or unconsciously, have an impact on the objectivity of that measurement. Campbell (1957) examined reactivity as one of many possible design flaws in experimental methodology. Stress, a component Campbell names as one that produces more intense reactive effects, touches on educational assessment systems at all levels, especially as stakes become high. He did not attach any single discipline to his findings but made clear 'social settings' were his subject area.⁶³ His early work lends itself well to many fields, including large-scale educational assessment.

⁶³ As echoed by Espeland and Sauder (2007, p.7) "Reactivity has a history and a literature that resonates across fields. The value of reframing work on the impact of new measurement technologies in terms of reactivity is that it offers a familiar language and a common focus that will encourage scholars to ask similar questions and make connections across a broader range of scholarship."

3.2.1 Other reactivity studies

Lerner and Tetlock (1999) examined how accountability measures can, in different cases, attenuate, amplify or have no discernible effect on workplace decision making and bias. They also relate that no amount of evaluation or accountability can *make* change happen – that depends on the abilities and skills found within the audited system.⁶⁴ Ferraro, Pfeffer and Sutton (2005) examined the self-fulfilling prophecy and how 'dominant assumptions' (like management practices, changing social norms, and the evolution of work language) colour the behaviour of actors in a system. They also note that accountability brings pressure to bear on the internal actors on a system, who, as a result, act most often in their own self-interest.

From far-ranging fields of study, reactivity effects are found. Matheson, Rogers, Katkutas and Dakos (2002, p.42) in therapeutic experimental studies state, "A test instrument is said to have reactivity or testing effect when the subject's experience of taking the test affects test performance." McCambridge and Kypri (2011) in a study of alcoholism wondered if self-monitoring assessments change behaviours and found that the influence was subtle unless the questions were primed to illicit reflection for change. Similarly, in a medical study on addictions, Hobbs, Walle and Hammersly (1979) found that a leading description of a self-assessment made it more likely to inspire positive reactivity. Rucsh, Menchetti, Crouch, Riva, Morgan and Agran (1984) studied the reactivity of employees to assessments when these were made obvious as compared to when they were covert. They found less reactivity apparent when observations were unknown to the employee.

3.2.2 Educational reactivity

More recently and also most comparable to this research, Espeland and Sauder (2007) developed ideas about educational reactivity in terms of the evaluation and ranking of U.S. law schools by an American national news magazine. The study evaluated how staff in law faculties changed their behaviours in order to accentuate their own school's positive qualities (i.e. those qualities considered positive in the magazine's assessment) in order to see ranking numbers increase. The authors make the case that the measurement itself creates the conditions that make the data less valuable: "Reactivity blurs the distinction between the act of measuring and its object. . ." ⁶⁵ They documented both the

⁶⁴ : . . no amount of increased effort can compensate for lack of knowledge about how to solve problems that require special training." (Lerner & Tetlock, 1999, p.263)

⁶⁵ Espeland & Sauder, 2007, p.3.

positive reactivity which came from more effective and purposeful work, and the negative which included the wilful gaming of data to manipulate overall scores. In the end, the purposeful gaming of the metrics led to the results having less predictive validity. This study was intended to answer their questions about accountability measures and the reactions they produce. Accountability and measurement, for Espeland and Sauder (2007, p.36), are not neutral or objective processes:

We have shown that public measures affect the distribution of resources, redefine statuses which can become reified and enduring, produce and reinforce inequality, and transform the language in which power presents and defends itself. . . . Accountability is routinely and uncritically invoked as an obvious public good, so it is especially important that scholars conduct empirically rigorous analyses.

It was expected that some of these same kinds of effects, positive and negative, might be employed by Canadian public school teachers who are held accountable by large-scale provincial tests. It is also thought that neutral reactivity effects (i.e. not using the data to inform instructional practices) might indicate that, as Espeland and Sauder (2007) noted, the measurement effect taints the data in the eyes of some making them appear less useful (for example, survey data include relatively high proportions of teachers who think that LSA results have **no** appropriate uses). In order to test this proposition, a unique model designed for determining the reactivity of teachers' instructional practices to education policy and external evaluation was used in this study.

In education, reactivity describes how teachers change their instructional practices for better or for worse as a result of external evaluation. Accountability policies differ in all provinces, and reactivity effects are also expected to be different. Witt, Noell, LaFleur & Mortenson (1997), for example, found that more monitoring and follow up on new educational practices led to more long-term and practical implementation. Nagy (2009) observed that some teachers will use the data as best they can, and change their practices in light of new information but others will seek out the apparent weaknesses in the testing scheme and work toward exploiting those. Using multiple measures of student achievement, a strategy long employed in classrooms, is one way to prevent the negative effects of high stakes, single test LSA systems (Koretz, 2002). It is commonly asserted that the mechanisms that currently drive educational reform are the LSA schemes (Ravitch,

2010; Koretz, 2009), but the connection from test to reactivity is not a clear one yet, as Cimbricz (2000) notes.⁶⁶

Improved instruction is the result of LSA policy when, assisted where needed by administrators, teachers take the results of mandated testing and use these data to inform their practice (Young, 2006; Volante, Cherubini & Drake, 2008). When LSA data are analyzed in ways that provide practical suggestions for improvement and growth (positive reactivity), this process can be seen as practical professional learning (Fullan, 2011; Stiggins, 2002; Ravitch, 2010). If the focus is the data themselves (i.e. scores on tests), policy goals seem to be more often corrupted and instruction suffers (Koretz, 2002).

Reactivity, where it occurs, equals change. Since reform is a stated goal of the drive for educational accountability, directing that change in a positive direction is pivotal. According to Corcoran & Goertz (1995) the real 'product' of the educational system is quality teaching, and this needs to be the primary focus of politicians, reformers, and educational professionals.⁶⁷ In cases of no change (neutral reactivity), it may be the result of professional staff not possessing the required skills to make a leap from simply administering a curriculum-based large-scale assessment to scoring, interpreting, analyzing and acting on the data from this same assessment (Shepard, Davidson & Bowman, 2011).

Figure 3.1: Summary of reactivity literature

Topic	Author(s)	Summary statement
Reactivity in the social sciences	Campbell, 1957	Study of flaws in social studies experiment designs which includes a section on reactive measures that taint any results with foreknowledge.

⁶⁶ "The studies reviewed suggest that while state testing does matter and influence what teachers say and do, so, too, to other things, such as teachers' knowledge of subject matter, their approaches to teaching, their views of learning, and the amalgam of experience and status they possess in the school organization. As a result, the influence state-mandated tests has (or not) on teachers and teaching would seem to depend on how teachers interpret state testing and use it to guide their action. . . How tests matter then is not always clear and simple." (Cimbricz, 2002, p.15)

⁶⁷ "We suggest that the defining "product" of the education system is high-quality instruction, which is central to the ability of the system to help all students reach high standards. . . Legislators and the informed public talk about reform as a means of improving teaching, which they see as the best means of helping students." (Corcoran & Goertz, 1995,p.27)

	Espeland & Sauder, 2007	The primary source of the author's reactivity model based on law school reporting of data for 'US News' national rankings. Results show cheating, stretching, and gaming.
	Hobbs, Walle & Hammersly, 1979	Study finds that informing people 'they will see change' primes reactivity effects (smoking, nail-biting).
	Hood, 2006	Paper on UK public service performance targets and how they led to gaming, analogous to Soviet era central planning issues.
	McCambridge & Kypri, 2011	A systematic review and meta analyses of brief alcohol intervention trials showing limited changes in behaviour based on surveys questions.
	Propper & Wilson, 2003	This study looks at 3 different cases of public service evaluation. There is an emphasis on 'mis-implementation' of policy and gaming.
	Rucsh, Menchetti, Crouch, Riva, Morgan & Agran, 1984	Investigates the reactivity effects of covert assessments as compared to overt ones.
	van Thiel & Leeuw, 2002	The performance paradox is the unintended consequences of NPM measures. Paper strong on background and educational examples.
	Webb, 2006	'Choreographed performance' is another term for negative reactivity, or gaming the metrics. Paper has an interesting take on performance to meet expectations.
Reactivity in education	Abrams, 2004	Compares survey results of all high stakes states to Florida results. They are much the same with lots of reactivity effects reported.
	Amrein, Berliner & Rideau, 2010	Categorizes cheating methods (negative reactivity) using a legal framework [pre-meditated, nuanced, unintentional/neglect].
	Au, 2007	Reactivity in implementation is examined using different terms: curricular control, formal control, and pedagogical control. This meta-analysis shows negative reactions are strongest.

	Booher-Jennings, 2005	Paper looking at the practical effects on teachers of high stakes accountability in Texas.
	Cimbricz, 2002	This study is a meta-analysis on teachers' beliefs. It shows there is reactivity and that tests driving instruction. It also shows curriculum narrowing, teaching to the test, and test-like assessment are common.
	Ehren & Swanborn, 2012	Dutch schools have evidence of test pool-shaping and not adhering to test administration rules. Examples of negative reactivity shown.
	Grant, 2000	Uses focus groups to look at New York state potential for instructional change with new testing scheme. There is not enough about actual strategies used here, but lots on impressions and attitudes.
	Haladyna, Bobbit Nolen & Haas, 1991	Looks at test score pollution and how high stakes policies drive poor practices. Includes a chart amended for use in Chapter 3 on ethical/unethical practices.
	Jacob & Levitt, 2003	A study on finding teacher cheating in Chicago. Findings are that it is extreme in 4-5% of classrooms but that moderate cheats go undetected.
	Jacob, 2004	LSA tests in Chicago are studied where increases appear in one school, but not in another. Evidence of gaming or strategic moves by schools.
	Koretz & Hamilton, 2003	A Boston study of teacher perceptions and reactions to high stakes tests done with surveys and interviews.
	Koretz & Jennings, 2010	Looks at testing policy issues using real world examples of reactions from teachers coaching and the problem of high stakes.
	Koretz, 2002	Digs into the practice of using tests to evaluate teacher performance. Focus on topics such as high stakes, score inflation, corrupt practices, what are 'meaningful' gains, and negative reactivity in general.

	Linn, 1998	Early but influential paper concludes that there is too much testing and too much emphasis on it. Examines the distinctive 'saw-tooth' pattern of test results which shows reactivity effects.
	Luna & Turner, 2001	Paper looks at the Massachusetts Comprehensive Assessment System tests using interviews. Concludes that high stakes leads to narrowing, test-like preparation, less choice and less creativity.
	Madaus, 1988	Paper concludes that high stakes tests are tied to many negative consequences and reactions.
	Nagy, 2000	Canadian study shows teachers do diagnosis with their own instruments. LSA issues are: test design, roll out, and minimal reactivity on external assessment.
	Noell, Witt, Gilberton, Ranier & Freeland, 1997	How follow-up with teachers improves implementation (not stakes, follow up and evaluation)
	Ravitch, 2010	Personal account of authors' changing opinions of charter schools, school choice, and school reform following from No Child Left Behind (NCLB) policy which included high stakes LSAs.
	Schorr, Firestone & Monfils, 2003	A New Jersey study looking at test design, PD opportunities, and stakes. In many cases, they find, teachers take on strategies in name only, the data have limited reactivity, and PD is test-based and ineffective.
	Volante, Cherubini & Drake, 2008	Examined principals' use of data. Administrators self-rated their skills, and differences were noted between elementary and secondary. Concludes that much PD needed at this level, as well.
	Young, 2006	Focus on leadership and other means to get data used namely alignment to curriculum and practices to facilitate data use.

3.3 Reactivity model

Campbell (1957) documented reactions to external assessment in social science domains, but LSA testing is relatively novel in Canadian public schools. As a result, there has been much debate about its effectiveness for improving student achievement, and much less on how teachers react to external evaluation. The fact that teachers are reactive to external assessment – especially when sanctions or rewards are written into policy (these are less commonly applied in Canada than elsewhere) – is not surprising.⁶⁸ Examining how they react, whether it is in ways that improve scores on test at the expense of the non-tested content, teaching to the test, or in ways that improve overall student learning is one of the purposes of this study.

Provincial assessment policies use a common perspective on the understanding of LSA scores and what they mean: better scores means that better learning has occurred in classrooms. Teachers are intended to clearly understand that their effectiveness as educators is measurable and can be quantified by increasing (or decreasing) scores on provincial tests. With a clear focus on this metric (improved student scores on LSAs), teachers then have the choice (assuming they have the requisite skills and knowledge to consciously make this choice) to take either high road or low road approaches to reach this end.

The high road approach will be defined as positive reactivity: those practices which are thought to be both ethical and broaden the number and variety of outcomes presented to students (see **Figure 3.2** below). This approach may be less applicable to improvements in any single test, but it would provide skills for students to be more effective in more situations than the low road approach. These strategies are well aligned with the Saskatchewan Teachers' Federation Code of Professional Competence and widely accepted pedagogical 'best practices' since they avoid the potential pitfalls of teaching to any specific test (Popham, 2001).

The low road approach is what has been defined as negative reactivity: those educational practices which are thought to be either unethical or reduce the number or variety of outcomes presented to students. The methods included in this approach would certainly be the most direct method to improve scores on a specific test (as such they are often referred to as 'teaching to the test'). As a result, 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 (Popham, 2001; Popham, 1999).

Therefore improved scores on provincial LSA tests, a common policy goal, might indicate something about students' learning and instruction, but they may

⁶⁸ Hanushek & Raymond, 2002, p.16-17: "The rewards and sanctions that many states have built into their accountability systems create the motivation for schools to change behavior."

not always highlight what is most important - that improved results are a sign of more or better learning (Linn, 1998; Koretz, 2002). It is also possible that higher scores, counter-intuitively, indicate less or more narrowed learning. Test scores and school rankings from LSAs do not necessarily provide a clear accounting of education even though this is what they were designed to do – this is Campbell's law in action.⁶⁹

These factors are the rationale for developing a unique reactivity model, that is intended to determine the level of teacher responsiveness (total reactivity) and also which of these teacher responses are constructive and effective (positive reactivity) and which are less constructive and less effective (negative reactivity) in helping students reach educational outcomes. All data were self-reported by teachers and were not observed first-hand. The proposed working model is built upon the terms of the Saskatchewan Teachers' Federation (STF) Code of Professional Competence. This code is in keeping with those seen across Canada and internationally in that it is has an aspirational component (it reaches for the very highest standards), an ethical component (it states how educators should behave to best protect students), and a conduct component (it sets professional standards on instructional practices; from: Nuland & Poisson, 2009).

Figure 3.2 shows the actual survey prompts used to determine types of reactivity employed in classrooms. The left is designated **positive reactivity**, which includes those reactions to LSA testing which are both ethical and provide a wide range of outcomes for students.

Figure 3.2: These are the survey items which asked respondent teachers to rate how commonly they employed these practices (if at all) in response to LSA results.

Positive Reactivity	Negative Reactivity
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.

⁶⁹ "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." Amrein & Berliner, 2002. p.5.

The right side of the chart includes **negative reactivity** practices which are reactions to LSA testing which are either unethical or provide diminished outcomes for students. The nation-wide survey was used to gauge both reactivity effects (the dependent variables) and the independent variables hypothesized to be correlated to reactivity. 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 were 'not at all', 'somewhat', or 'a great deal.' They were not identified by name as positive or negative reactivity.

The positive reactivity side of this chart also adheres to the standards set out in the STF Code: (a) practices that exclude content or exclude opportunities for play are avoided; (b) the expectation of knowledge and expertise related to the curriculum; (c) instructional methods are to be varied and appropriate; (d) assessment and evaluation are to be done in accordance with ministry and professional expectations; and (e) a teacher is expected to be reflective about all of these inter-related practices to improve one's educational practice. The five items on the positive list are within the boundaries of these conduct guidelines. These practices are ethical and provide strong educational outcomes for students. They are also very much in line with the aspects of 'professional accountability' as spelled out by Møller (2008).⁷⁰

On the right side negative reactivity practices are enumerated. (1) Curriculum narrowing and teaching to the test violate the 'support of the whole child' philosophy (by skipping 'less important' content, that may be both fun and interesting to the student) and go against the 'professional level of knowledge about the curriculum' by removing parts of it from the instructional domain. This practice leads to the assessment domain effectively becoming the taught curriculum. (2) Teaching test-taking skills is a specific form of deviation from the curriculum. It may improve LSA results, but it is nevertheless not a 'leverage' skill, and is not included in any provincial curriculum documents known to the researcher. (3) Test-like classroom assessments certainly prepare students for the provincial tests but delimit a student's means of expression and limit the teacher to specific curriculum domains. Provincial assessments are written to be marked in a

⁷⁰ "There is also a professional accountability, where a person's commitment to a community of professionals makes him/her perceive a duty to adhere to the standards of the profession. This is about teaching as a moral endeavour. Codes of ethics have for instance become a familiar part of the rhetoric of professional control of the work in schools, even though the influence of these codes is uncertain. Professional accountability implies that teachers acquire and apply the knowledge and skills needed for successful practice. In addition, it involves the norms of putting the needs of the students at the centre of their work, collaborating and sharing of knowledge, and a commitment to the improvement of practice." (Møller, 2008, p. 40)

more or less standardized way, and not to open the door to problematic, subjective marking which come from expressiveness and creativity. Classroom assessments that mirror LSAs have these same restrictions. This violates the third STF Code item: using a mixed repertoire of instructional strategies. (4) Less time for non-core subjects is a similar negative response since removing a student from band or gym class is not in keeping with the whole child philosophy or the intent of the provincial curricula. (5) In some jurisdictions, the exams are so similar year on year that reviewing old exams has great potential to improve overall LSA performance. Test similarity is one reason that provinces are often unwilling to share specific test responses with teachers directly, thus making the results they receive less useful for improving instruction. These past exams, well-written though they may be, set finite boundaries on what is taught to the exclusion of material in the curriculum that might otherwise be explored.

No questions in the survey asked directly about the most egregious practices discussed in the reactivity literature, as it was expected that very few teachers would admit to practices such as helping student during exams, changing test answers, or providing students with actual exam questions in advance. It is also considered taboo to admit a policy of actively excluding low-performing students, for whatever reason or in whatever manner, as it violates professional responsibility in regard to assessments. In these cases, there is no possible defence that instruction or learning has been improved. These topics were touched upon in some respondent comments and were explored with interview subjects.

Examples of these kinds of positive and negative reactivity are prevalent in the literature. This study takes this work as a starting point to further examine which Canadian assessment policies produce the most positive outcomes.

3.4 Reactivity survey results

The following provincial and national data relate to **reactivity** which is the dependent variable of this study, and will influence all analyses in the dissertation. The data were collected from individual respondents in all ten Canadian provinces and have been aggregated into provincial and national data sets. Specific details from the data collection and variables are found in the chapter-ending section and include **Figures 3.4 – 3.15**. The following dependent variables (DVs) are examined:

- DV1 (dependent variable 1) - positive reactivity
- DV2 - negative reactivity
- DV3 - total reactivity (all reactivity effects, positive and negative)
- and DV4, net reactivity (positive and negative added to allow cancelling)

Reactivity is defined as any change in behaviour resulting from an external assessment being used to measure performance. As provinces all give provincial assessments, there is bound to be some reactivity in the teaching populations for these jurisdictions. How much total reactivity, and what kinds of reactivity (positive, negative or neutral – which is itself a low total reactivity score) will depend on the personal and policy factors that are examined in this dissertation. To begin, then, the amount and types of reactivity prevalent in each provincial sample of survey respondents will be examined.

Five questions about positive reactivity practices were asked and followed by five questions about negative reactivity practices. For each one, a strategy was given and the teachers were asked whether they had used that strategy 'a great deal,' 'somewhat,' or 'not at all.' The positive reactivity responses were given numeric values (0, +0.5, or +1) and were added to give each respondent a score from 0 to 5. The negative reactivity responses were also given numeric values (0, -0.5, or -1) and added together to give each respondent a score from 0 to -5. Total reactivity is the positive scores added to the absolute value of negative scores which are examined to show how much reactivity of either type is evident (on a scale of 0 to 10). Net reactivity allows for the cancelling of positive and negative values attached to the practices to see which side of the spectrum has greater sway in each province and nationally (the scale runs from -5 to +5).

Cronbach's alpha was used as an index of reliability for both positive and negative reactivity scales. An alpha score of 0.7 or more indicates adequate or better 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 positive reactivity items was 0.62 and for negative reactivity items was 0.76. The correlations between the items of the positive scale were within a close range so that it was not possible to improve the alpha by dropping items. These data do not, in the author's mind, undercut the value of the analyses that follow. These values are subject to interpretation and that shall be left to the reader except to say that it has been noted (Tavakol & Dennick, 2011) that having fewer values in aggregated scales may result lower alpha scores. Both of these scales consisted of 5 survey items.

Note that the following analysis is based on the data shown in the chapter-ending charts and tables section (**section 3.6**), and that the values given to survey responses for regression purposes can be seen in **Annex 2**.

3.4.1 National data

- DV1 - positive reactivity (details are found in **Figures 3.4 - 3.6**)

The self-reported national survey data for positive reactivity effects show a very evenly distributed curve of positive reactivity effects, although somewhat under-

dispersed (likely a result of the small range of choices available). The least reactive respondents fall into the aggregated 'neutral positive' category (so named since less reactivity is present) and followed in turn by 'moderate positive' and 'strong positive' groups. The moderate group makes up 47% of all respondents ($n = 418$), with neutral positive and strong positive nearly equal (26% and 27% respectively).

- DV2 - negative reactivity (details are found in **Figures 3.7 - 3.9**)

Negative reactivity effects are more unbalanced and show deviations from the normal distribution. The mean is higher (there is more negative reactivity than positive), and the curve is also skewed left indicating high proportions of negative reactivity effects. Aggregated data show 'strong negative' reactivity as the most common response (44%) followed by 'moderate negative' (43%) and 'neutral negative' (14%).

- DV3 - total reactivity (details are found in **Figures 3.10 – 3.12**)

Total reactivity adds the absolute values of positive and negative reactivity scores, thus all values over 5 indicate the use of both positive and negative reactivity practices. The mean is above 5 and indicates that the mixing of positive and negative effects is more than common, it is routine. Aggregated data show this well with 'strong reactivity' having the highest proportional response level (39%) followed by 'moderate reactivity' (29%), 'very strong reactivity' (20%), and final 'neutral reactivity' (13%).

- DV4 - net reactivity (details are found in **Figures 3.13 – 3.15**)

Net reactivity allows us to see the balance between positive and negative effects. It has been seen that national data indicate more negative responses than positive ones, and as the negative numerical values cancel out the positive values, the extent of this trend becomes apparent. By far and away the largest proportion of respondents fall into the neutral category (62%) but the tails on the negative side, while small, are larger than the tails on the positive side of the curve. 'Moderate negative' reactivity is recorded by 28% of respondents ('moderate positive' is 8%) and 'strong negative' effects make up 3% of respondents ('strong positive' is 1%).

- Summary of reactivity effects

All the data collected nationally on reactivity show that teachers are quite reactive to LSA data. The total reactivity scores are the best indication of this fact with the mean score here being higher than 5 (the halfway point of the 10 point scale). A less reactive group of respondents would have had a lower mean score, and perhaps might also have opted for one type of reactivity (positive or negative) more than the other. Scores above 5 also show that positive and negative effects co-exist with most teachers (since a score of 5 is the maximum possible value from either positive effects alone or negative effects alone). National and provincial average scores can be found in **Table 3.3** below.

In terms of the reactivity effects that are most commonly practiced, the survey data make clear that there are more negative reactive practices reported

than positive ones. Negative reactivity indicates practices that do not have the leverage to transfer over the subject domain completely, but rather focus on the skills and content most-tested and/or on the tips and tricks that can help a student pass a particular test (Koretz & Jennings, 2010). Some of these practices are unethical, according to Haladyna, Bobbit Nolen and Haas (1991), and some narrow the number and scope of curriculum outcomes covered. According to the STF Code of Professional Conduct, these same practices are less effective and un-professional. The distinction to be made between the 5 negative options given on the survey and the 5 positive options does not seem to have been an obvious one for any of the respondents, and honest responses to interview questions have confirmed this apparent ambiguity.

Positive reactivity effects were also quite common, though, and in some provinces, they nearly balance with the negative practices. The 5 choices given to respondents here do have a common thread in that they put the emphasis for extra work at least in part on the classroom teacher. Many respondents obviously do not balk at this aspect of positive reactivity effects and use them frequently.

Net reactivity scores are another means of showing the balance between positive and negative, and these indicate the scales are definitely tipped toward the negative. Since it has been seen that total reactivity scores are high as well, the national sample of 418 teachers who give large-scale provincial assessments have provided the data necessary to examine these effects further and to try (in upcoming chapters) to untangle which policy factors promote reactivity in general, and which factors promote the more ethical and effective positive reactive practices.

Table 3.3: Ranking Canadian provinces based on reactivity effects. The provinces are rated below based on the four different metrics of reactivity from survey data. All respondents' scores were totalled, and divided by *n* to provide an average score. Raw data can be seen in **Figures 3.4 – 3.9**.

POSITIVE REACTIVITY			NEGATIVE REACTIVITY			TOTAL REACTIVITY			NET REACTIVITY		
	Avg. score	Rank		Avg. score	Rank		Avg. score	Rank		Avg. score	Rank
AB	2.95	1	ON	-3.73	1	AB	6.39	1	NS	0.21	1
NB	2.84	2	QC	-3.52	2	QC	6.29	2	PEI	-0.07	2
PEI	2.82	3	NL	-3.53	3	NL	6.26	3	SK	-0.21	3
QC	2.78	4	AB	-3.45	4	NB	6.03	4	MB	-0.30	4
NL	2.72	5	NB	-3.19	5	ON	5.99	5	NB	-0.35	5
MB	2.41	6	BC	-3.09	6	PEI	5.72	6	AB	-0.50	6
NS	2.34	7	PEI	-2.90	7	MB	5.11	7	QC	-0.74	7
ON	2.26	8	MB	-2.70	8	BC	4.90	8	NL	-0.81	8
SK	2.17	9	SK	-2.38	9	SK	4.55	9	BC	-1.28	9
BC	1.81	10	NS	-2.14	10	NS	4.48	10	ON	-1.46	10
CANADA	2.50			-3.04			5.53			-0.54	

3.4.2 Provincial results

Alberta

- DV1 - positive reactivity (details are found in **Figures 3.4 - 3.6**)

In terms of the average positive effects score from respondents, Alberta is the highest rated province (2.95 in Alberta whereas nationally the average is 2.50). Alberta has 'strong positive' reactivity (an aggregation of the highest reactivity scores) well above the national average (35% to 27%) and 'moderate positive' slightly above. Obviously, the 'neutral positive' (positive reactivity at low levels) is much lower (15% to 26%). (For specifics, see the 'Charts and tables' section at the end of the chapter)

- DV2 - negative reactivity (details are found in **Figures 3.7 - 3.9**)

Alberta also has 'strong negative' reactivity well above the national average (59% to 44%), which is itself higher than that for 'strong positive'. Both moderate and neutral negative numbers are lower than national figures. The average is fourth lowest of all provinces (-3.45 here, -3.04 nationally).

- DV3 - total reactivity (details are found in **Figures 3.10 – 3.12**)

The numbers from above clearly indicate that total reactivity is strong in Alberta. The average total reactivity score is 6.39 (top rated in this category) compared to 5.53 nationally. 'Very strong' reactivity (33%) is well above the national average (20%), and the fall in 'neutral' reactivity (4% compared to 13% nationally) numbers makes up almost all of this difference.

- DV4 - net reactivity (details are found in **Figures 3.13 – 3.15**)

To look at the balance of positive and negative, there is more negative reactivity evident from Alberta teachers than positive. Their net reactivity effects results are almost identical to the national averages (an average score of -0.50, while nationally it is -0.54).

- Summary of reactivity effects

There is definitely a large amount of reactivity evident in the responses from Alberta teachers to survey questions. With high numbers in both positive and negative, there seems to be no clear distinction made by teachers between positive reactivity and negative reactivity. Interview responses have borne out the conclusion that teachers and administrators in this province have a hard time dismissing negative reactivity effects because the tests have so much value in the system (this is true of only four other provinces, and is also a variable examined in Chapter 7). In the context of negative reactivity, test scores do not necessarily mean that students are learning more or learning in more depth, but they are made more familiar with the expectations of the test through test-driven instructional choices, test-like assessment practices, focusing even non-core subject teachers toward test

score improvement, and system-wide recognition that LSAs are the only metric common to schools in the province.

I don't think in and of themselves [provincial tests] do [improve teaching] but they have the potential to. You know, I would answer with a question and say, in the absence of provincial testing how would you know how to improve your teaching if you don't have an objective measure?

- AB, High school Science teacher, male

They [my class] learned a lot more than they would have if we had had the PATs (large-scale tests) this year because I could actually teach the curriculum and the students instead of trying to worry about teaching them how to fill in Scan-tron (machine-scoring) bubbles and how to write a copy-cat story.

- AB, Elementary English teacher, female

In fact, one should not dismiss the fact that many positive reactive effects are also evident. Alberta has more positive reactivity than any other province. Alberta teachers and administrators have indicated that the LSAs have been effective at focusing instruction on the curriculum; allowing teachers to collaborate to work with the data; guiding additional resources toward improving the scores (specifically of those students 'on the bubble'); and giving schools a means of comparing 'apples to apples' when looking at student performance. The main specific difficulty cited in this regard was the lack of detailed information provided by the ministry which, if it were provided, would allow teachers to make more sense of the results data and act accordingly.

I get a rough idea of how my students performed on each of the general learner expectations. . . Then I get a rough idea of how I should maybe change teaching practices or maybe change my delivery to better match up with the students. But again without seeing the questions it is kind of difficult to see, you know. Is it something that I have taught? Is it something . . . I taught badly? Is it something I need to expand on? I don't really know at the end. I just know that my students didn't perform well, but I'm not sure exactly why. **- AB, High school Science teacher, male**

The results refer only to numbered curriculum outcomes which are very broad in nature and thus cannot identify a particular area of

concern. Because we cannot see the test, it is difficult to have complete understanding of the results.

– **Anonymous survey comment**

The Alberta LSAs are, by definition from the Education Ministry, expected to assist schools in monitoring and improving learning. They are also set the task of revealing areas of strength and weakness in achievement. The positive effects noted above (collaboration, focused instruction, additional resources) are steps that need to be taken in greater measure for these policy purposes to be fulfilled.

British Columbia

- DV1 - positive reactivity

British Columbia respondents showed much less strong positive reactivity (7.7%) than the national sample (27.0%). The average score is the lowest across Canada at 1.81 (2.50 being the average nationally). Moderate positive was somewhat less than national figures, but the main difference is the large number of neutral positive responses (51.3%).

- DV2 - negative reactivity

Respondents in British Columbia showed a high level of strong negative reactivity (53.8%). The level of neutral negative was in line with national averages, but comparatively more moderate negative was also evident here. The average negative reactivity score is -3.09, very close to the national -3.04.

- DV3 - total reactivity

In total reactivity effects, British Columbia is very close to national scores with somewhat less strong reactivity and somewhat more neutral. The provincial average (4.90) is lower than the national average (5.53). It has been discussed that most of this total comes from negative effects.

- DV4 - net reactivity

The balance tips toward negative effects in this metric as negative responses more than cancel out positive reactive responses. Neutral and moderate negative effects are strongest at 48.7% each. There are no strong positive effects indicated but 2.6% strong negative. The average net score in BC is -1.28: the second lowest nationally and well below the national average of -0.54.

- Summary of reactivity effects

British Columbia policy makers should be concerned with how teachers are using their provincial assessment data. There are significantly more negative effects in evidence from this survey than positive effects. We have seen that positive reactivity effects provide students with more inclusive educational experiences whereas the negative effects focus on test-score improvement as an end in itself. The data also show that British Columbia teachers are very close to as

reactive (in total) as other teachers in Canada (see **Table 3.3** above). The difference is that BC teachers have a large portion of this total reactivity coming from the negative side of the spectrum.

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. . . I give them a practice. We go into the computers and I show them how to manipulate the programs so they know how to work through the testing and that is it.

– **BC, Elementary homeroom teacher, female**

[Teaching to the test] is an inappropriate outcome that comes from the exam but when the teacher chooses to do it from the best intentions, to help kids out, I can't say it is inappropriate. I can say it is a negative consequence of having a provincial exam that it is making people feel they need to do that. - **BC, Division staff, male**

Whether they are not granted free and frequent access to the kinds of support that would motivate more positive effects is a subject for a more focused study.

The position that the BC Teachers' Federation holds about standardized testing is an extreme one, and does not favour the use of these tests in BC schools. The teachers in this jurisdiction are also working in a politically charged time, and this was certainly true when the survey was written. I would suggest, though, that negative reactivity is not a response suited to a dislike or distrust of the assessment instruments – that response would be to ignore the assessments as much as possible (neutral effects). Nor are negative effects a likely result of labour action between the Federation and the ministry – again, to dismiss the results as irrelevant (neutral reactivity) would be a more overt act of defiance. Negative reactivity indicates to the researcher that somewhere between the drafting and implementation of assessment policy British Columbia teachers have made the conscious or unconscious choice that the LSAs in this province do make a difference somehow, but they are quick to show that they do not appreciate the manner in which assessment is both done and reported.

That's it, the Fraser Institute . . . published results in the newspaper, and says these are the top schools and these schools aren't doing very well. Our particular school, we are in an area with high socio-economic needs. And there is another elementary school in my

community and it caters, like a lot of the doctors and research scientists who live in that area, so there, their school usually does really well. But if you look at it, our kids are doing just as well.

- **BC, Elementary homeroom teacher, female**

Our teacher union, the BCTF (the British Columbia Teachers' Federation), has taken opposition to that [use of the data in classrooms] and said that this is not something that is appropriate and therefore you have a lot of resistance from teachers in general around not just administering it but using the results for anything in particular. - **BC, Division staff, male**

For some parents who get concerned, I think then we can say, you know, this [provincial testing] is one of the reasons why we are out on job action or this is why we are calling for more support.

- **BC, Elementary homeroom teacher, female**

If FSAs are to be 'a resource to support teaching and learning' there needs to be a more consistent approach to the use of the results data by teachers in the province. Reactivity needs to be harnessed and focused on strategies that will meet ministry expectations to 'provide support for learners.' If the test score takes precedence over learning, this objective cannot be reached.

I would say there are other people who teach who agree this is a great thing. You know, why wouldn't we have a standard sitting there that we can teach towards? It is clear that this is what people feel is important and we're going to make sure all the kids know this stuff. And the easiest format is it's all sitting right there in an exam. . . You get both approaches, but it is a reality, that there are people teaching to provincial exams who don't think it is the best thing to do.

- **BC, Division staff, male**

Manitoba

- DV1 - positive reactivity

Manitoba respondents differ very little from the national average for this metric, with only slightly more neutral positive (30% compared to 26%), and slightly less strong positive reactivity (22% compared to 27%). The average score is 2.41 compared to the national average of 2.50.

- DV2 - negative reactivity

Manitoba has less strong negative reactivity (35% compared to 44%) and more neutral negative (22% compared to 16%) as compared to national averages. The average in this metric is -2.70, higher than the national average of -3.04.

- DV3 - total reactivity

Since positive and negative reactivity tends towards the neutral range, the total neutral effects overall in Manitoba teachers are more pronounced than the national average (24% compared to 13%). Moderate and strong total reactivity effects are much the same as the national average, with the difference coming from very strong reactivity effects (much lower at 11% in Manitoba and 20% nationally). The average respondent score is 5.11 compared to a 5.53 national average.

- DV4 - net reactivity

Net reactivity is very much in line with national average, but somewhat more moderate negative effects (24%) in the net calculation than moderate positive effects (11%). The average score (-0.30) is closer to the positive range than the national average (-0.54).

- Summary of reactivity effects

Manitoba is not the least reactive province, but it is in the running for that title behind British Columbia, Saskatchewan and Nova Scotia (Table 3.3 above). Between neutral and moderate total effects percentages, 51% of Manitoba respondents indicated that LSAs had very little effect on their teaching practices. Of course this means 49% indicated strong effects, but these were just about equally split between positive and negative sides of the scale, so even this result is not encouraging. Like Alberta and British Columbia, in Manitoba there is more negative reactivity than positive, especially in the 'strong effects' category.

It has just become such standard practice that at this point I think it's . . . I don't know. I think it is something we obviously need to look at, you know. Is it necessary, but there is no conversation.

- MB, Elementary school principal, female

The nature of Manitoba LSAs in elementary and middle years grades is quite different from paper and pencil test given in all other provinces. This may be one reason that neutral reactivity is so strong. Neutral reactivity is certainly not as pernicious as the negative variety, but it is not without its costs: if the time and expense used for administering and analyzing provincial LSA assessments is not providing any impetus for instructional change in the classrooms of Manitoba, then the effectiveness of the assessment program should be questioned. It is also true that the Ministry of Education has instructional improvement ("informing instructional planning and helping to determine the need for . . . student specific

intervention”⁷¹) as an explicit goal of the program, so implementation does not seem to have the necessary fidelity to carry the policy intentions all the way through to action by front-line educators.

It is not like we are shipped a bunch of tests and the kids all complete the same thing. So every teacher would assess it and administer different things to collect the data differently. And I know that even within our school, like my standard for what I give for a certain grade of for a certain 'meeting' [expectations grade] or 'approaching' [expectations grade] is very much different from the teacher who teaches across the hall from me who thinks that everyone will get the highest score because they have shown that they tried. – **MB, Middle years Math teacher, female**

The policy and the assessment instrument were designed to facilitate positive reactivity effects, but this is not apparent from respondents.

New Brunswick

- DV1 - positive reactivity (details are found in **Figures 3.4 - 3.6**)

Strong positive reactivity is 11% higher than the national average in New Brunswick, and the difference comes just about in full from the proportional neutral positive effects (12% lower than the national average). The average score (2.84) is the second highest nationally (average 2.50).

- DV2 - negative reactivity (details are found in **Figures 3.7 - 3.9**)

Strong negative reactivity is just about on par with the national average, but moderate negative effects are 8% higher, the difference coming out of neutral negative effects. The average score is middling at -3.19, where the national average is -3.04.

- DV3 - total reactivity (details are found in **Figures 3.10 – 3.12**)

As follows from the relatively strong positive and negative effects above, New Brunswick is quite reactive and has 72.7% of teachers in the strong or very strong total reactivity categories (compared to only 58.6% nationally in the same two groups). The average rating from respondents is 6.03, substantially higher than the national average (5.53).

- DV4 - net reactivity (details are found in **Figures 3.13 – 3.15**)

The net effects show strong balancing of positive and negative practices, which means there is a 66% proportion of respondents in the neutral category. Only Prince Edward Island and Saskatchewan have higher 'net neutral' effects than this.

⁷¹ Manitoba Education, retrieved Aug. 9, 2014 from: <http://www.edu.gov.mb.ca/k12/assess/>

The larger proportion of negative effects overall is telling in that 26% of respondents are neutral negative compared to only 9% neutral positive. New Brunswick has a middling average score of -0.35 compared to -0.54 nationally.

- Summary of reactivity effects

Total reactivity in New Brunswick is quite high indicating that the policy factors here are certainly having at least some of the desired effect. Only Alberta, Newfoundland and Labrador, Québec teachers have a higher proportion of total reactivity than New Brunswick (**Table 3.3** above). It seems that filing and forgetting the LSA results is just not a common practice in this province.

Our focus is definitely on trying to bring up the areas that are weak. - **NB, Middle years homeroom teacher, female**

I think [provincial tests] can [improve teaching]. Absolutely. I'm going to say not because of the accountability factor where teachers may teach to the test. . . But I think, what I would hope is, by having provincial assessments, that teachers take the time to reflect on their strengths and their weaknesses and that is key - that they have to reflect themselves. . . And then you find the time to work collaboratively with a colleague and then you are growing as an individual which means the next class that you teach, it is going to benefit from that. So these provincial assessments do more than just hold, I don't like the word accountability, but they allow for professional growth as a teacher. . . Now if you are a teacher who puts their feet up on the desk and doesn't give a rat's ass about any of those things, well then, I'm to discover that anyway. - **NB, High school principal, male**

The net effects are not as promising, though, with net neutral proportions being very high, followed by moderate negative effects.

All provinces examined thus far have shown more negative reactivity than positive. New Brunswick teachers are more reactive than the three provinces above, but no more discerning about the variety of reactivity that is employed. There seems to be a large amount of uncertainty regarding those instructional practices related to large-scale assessment that provide the best (or the most) educational outcomes for students.

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 . . . what types of response will they be looking for. Now you could say . . . that is teaching to the test but it isn't because we have no idea what the content is like. - **NB, High school principal, male**

If LSAs are going to live up to the lofty purposes for which they were designed, there must be more accurate and consistent information given to teachers about what practices are best in this regard. My conversations with educators go further to clarify the kind of communication breakdowns that can occur between test policy and practice:

We're talking about an individualized learning plan for students nowadays, and we have had a couple of students who are on special learning plans, or individualized learning plans and they have still had to write a provincial assessment. And that is just really ridiculous because we already know that they are far, far below grade level and that they are never going to meet that level of outcome and yet the students are still put through that level of pressure. - **NB, Middle years homeroom teacher, female**

The Department of Education and Early Childhood Development in New Brunswick has a great deal of information available to educators, parents and students justifying the large-scale assessments they use, and how they are intended to improve teaching and learning. To 'tailor instruction' for learners takes assessment literacy, analysis skills, and time to do the work. While one cannot make an omelette without breaking some eggs, and it seems that in many provinces, the necessary skills for the tasks at hand are missing, spotty, or under-utilized.

Newfoundland and Labrador

- DV1 - positive reactivity

Newfoundland and Labrador teachers had similar proportions of strong positive reactivity as the national sample, but moderate positive was 8% higher, this coming at the expense of neutral positive. The average score is 2.72, higher than the national average of 2.50.

- DV2 - negative reactivity

Neutral negative reactivity was not reported at all from respondents in this province. Moderate negative was just over 1% lower than national figures and strong negative was 15% higher. The average score is -3.53, lower than the -3.04 national average.

- DV3 - total reactivity

With very low proportions of neutral effects in both positive and negative categories, total reactivity in Newfoundland and Labrador was very high. 55% of respondents showed strong total reactivity and 24% showed very strong effects. These levels are respectively 16% and 4% above national levels. With 79.3% of respondents in the strong or very strong groups, Newfoundland and Labrador is tied with Québec as the most 'highly reactive' province in Canada. The average score of 6.26 is the third highest rating for a province and significantly higher than the national average (5.53).

- DV4 - net reactivity

Looking at aggregated net effects, there is no reporting of either moderate or strong positive reactivity. All respondents fall into the neutral (59%), moderate negative (38%) or strong negative (3%) categories. The average on this metric (-0.81) is lower than the national average (-0.54).

- Summary of reactivity effects

In some respects, reactivity to testing in Newfoundland shows success: there is no doubt that educators in this province take the provincial assessment seriously and use them, for good or for bad, to inform their instruction. Almost four out of five have strongly reactive effects apparent in their responses to the survey. Yet part of the assessment policy in Newfoundland and Labrador includes (like Alberta) high stakes grade 12 exams for students. These are important marks for gaining university entrance, and to qualify for scholarships as well. So in speaking with teachers from high schools, one cannot help but think that these teachers have to, in the current circumstances, do what they can to prepare students to do well on these tests:

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

- NL, High school Science teacher, male

The supposed success of the LSA policy assumed above does not carry through when the conversation switches to best practices and to those skills that have leverage beyond the exit exams of high school. Negative reactivity effects are the surest methods to see test scores improve, but they don't have any momentum beyond these very specific tasks. Knowing more about the test format, about commonly tested domains, and about how to answer with a provincial scoring

rubric in mind are useful to do well on these tests – and it is important (too important, possibly) that students do achieve the highest marks they can on these exams. Perhaps the pressure applied here is not able to produce the educational outcomes the policy proposes and expects.

But I wonder about that sometimes. It bothers me a lot, because, you know, you can start having an argument that I'm teaching to rote memory, in a way, to the test as opposed to honest-to-god chemistry. . . . But then I've got weak kids, and I've got to get them through too. You know your 50s, 60s students they want to pass . . . and sometimes they need that A, B, C.

- **NL, High school Science teacher, male**

When the stated goal of the ministry is 'improving student achievement' there are high road and low road approaches to get there. I do not fault teachers for understanding that if this is the goal, then there are certain negative reactive practices that show a clear path to higher test scores, and thus better student achievement based on the sole universal metric in the province. The question that hangs over this action-reaction dynamic is whether higher scores on domain-limited, one-time, subject-specific assessments really do indicate higher levels of achievement. In lieu of another system or measure, the working assumption of teachers, schools and boards, must be that it does.

Nova Scotia

- DV1 - positive reactivity

Moderate positive reactivity is the most prevalent in Nova Scotia respondents (50%) followed by neutral positive (27%) and strong positive (23%). These are very near to national averages, however the average score here (2.34) is somewhat lower than the national average (2.50).

- DV2 - negative reactivity

For negative reactivity effects, the neutral option is more than double the level of strong negative (35% to 17%). Moderate negative is the highest rated effect (48%). Nova Scotia teachers depart from the national average significantly with much fewer strong negative effects and commensurately large proportions in the neutral negative. The highest provincial average is recorded here at -2.18 compared to the national average of -3.04, indicating quite low negative reactivity.

- DV3 - total reactivity

The tendency in Nova Scotia is toward neutral effects, and this means total reactivity scores are lower. There are more respondents in the neutral (14% more) and moderate groups (2% more) than is true nationally, and fewer in the strong

(4% less) and very strong (14% less) groupings. They have the lowest average (4.48) for this metric in the nation (national average 5.53).

- DV4 - net reactivity

Net reactivity tends toward the mean as well, which is neutral effects (60%). Yet this is the first province where net reactivity tends to the positive side – there is more moderate positive (23%) than moderate negative (13%). The average score is positive (0.21) when all other provinces have negative averages and the national average is -0.54.

- Summary of reactivity effects

As with most provinces, there are some positive signs in these data from Nova Scotia, and there are things that bear further scrutiny. The best news is that Nova Scotia teachers seem to use positive reactivity strategies more commonly than negatively reactive ones. Nova Scotia is the first province to be examined where positive reactivity outweighs negative reactivity. The bad news is that it will also be the last. There can be no doubt that as more provinces are seen where negative reactivity effects are predominant as a means of addressing LSA results, what appears is a major disconnect between assessment policy intentions and the follow through on the ground. This needs to be addressed by all education ministries and not just by proactive individuals, schools or school boards.

The other thing I think, for us in the . . . school board that has I think been a positive outcome from the assessment process is that we have actually looked at the results and used them for instructional purposes and for developing interventions. . . So there was direct correlation between those results and action that has been put in place. - NS, Division staff, female

The more troubling provincial news is that Nova Scotia teachers are not particularly reactive to LSA data. Only Saskatchewan has more respondents in the neutral or moderate reactivity groupings for total reactivity, and the provincial average on total reactivity is the lowest in the country. Part of this result can be ascribed to the fact that Nova Scotia teachers do not readily employ negative reactive practices, and that takes half of the options off of the table. Yet there are many teachers who do use negative practices, if less frequently than positive ones, and it is worth noting that the cut-off score for 'moderate' in the total reactivity category was a score of 5, with a full point given to the regular use of any of the ten strategies listed (less than regular use was a half point). So Nova Scotia teachers, whose moderate and strong negative reactivity proportions total 65% of respondents, do not show a great deal of reactivity in general. There appears to be a lack of understanding as to what is expected from the test instruments, and some reactions noted have been of the less effective variety.

I'm not sure what you mean by "act on assessment results."

- Anonymous survey comment

I think some do [change classroom assessment to match provincial assessments], and I think that is sort of the 'pros and cons' piece. It's that sometimes we don't want people just doing assessments modelled after provincial assessments because that's a very particular type of test and very limiting. . . So I think people try to be really careful about that and certainly our principals are really aware of that.

- NS, Division staff, female

It is worth asking whether this lack of reactivity begins at the policy level. It was noted in the introductory chapter that the Evaluation Services branch of Nova Scotia's Department of Education and Early Childhood Development has a long list of goals and objectives. Some of these relate to policy-level decisions ('determining how effectively the curriculum is delivered'), while some relate directly to teachers in their classrooms ('assisting students to meet outcomes'). Such a wide range of activities are certainly difficult to align, and using one assessment instrument for all of them, namely provincial LSAs, is not ideal.

I don't think the assessment by itself improves teaching. I think it depends, first of all, on how the assessment is constructed and, number two, what you do with the results of those assessments and if they impact on the actual day-to-day planning and interventions. And if you can take those and say, "Yes, it has made a difference and we were able to get the resources that we needed. We were able to do the things we wanted to do." Then, I think, we can say, "Yes, they were useful." - NS, Division staff, female

So the distinction must be made clear, in that diagnostic testing, which provides detailed single-student information on specific outcomes and skills, takes time to administer, to interpret, and to use in practice. Program evaluation does not need this level of specificity, nor the census-style delivery. Using different instruments for these tasks may be a better way forward for Nova Scotia and other jurisdictions that expect provincial tests to be all things to all stakeholders.

Ontario

- DV1 - positive reactivity (details are found in **Figures 3.4 - 3.6**)

Ontario has most uniform set of responses for positive reactivity with only 13% separating the lowest-rated (strong positive, 26%) and the highest-rated (moderate

positive, 39%) responses. The average score in Ontario is 2.26, lower than the national average of 2.50.

- DV2 - negative reactivity (details are found in **Figures 3.7 - 3.9**)

Negative reactivity is a world away from the positive results just examined: only 4% of Ontario respondents showed neutral negative effects while 67% indicated strong negative reactivity. This is the highest national figure for strong negative effects, and 23% higher than the national average. The average score in Ontario is -3.73, the lowest nationally (the national average is -3.04).

- DV3 - total reactivity (details are found in **Figures 3.10 – 3.12**)

Ontario is in the middle of the field and close to national averages when looking at total reactivity. The strong negative effects noted above are offset by the tepid positive effects.

- DV4 - net reactivity (details are found in **Figures 3.13 – 3.15**)

Net effects show a quite substantial bend toward negative reactivity. With cancelling in effect, there are no reportable moderate or strong positive effects. Nearly half of respondents fall into the neutral group (49%) with more than half of teachers rating moderate or strong negative effects. The average net score on Ontario (-1.46) is the lowest nationally (average -0.54).

- Summary of reactivity effects

Ontario's EQAO office is often touted as the most comprehensive and professional public sector educational testing body devised in a Canadian province. This may well be true, but there is also a fairly notable reactivity issue in Ontario. Whether EQAO's strong message is not getting through to teachers clearly or the message itself is not very clear, it is apparent from these data that Ontario is the third least positively reactive province, and the most negatively reactive province. No other Canadian jurisdiction has reactivity effects where both sides, in effect, come up tails. Manitoba, for example has a low proportion of positive reactivity, but also a low proportion of negative reactivity – teachers here are not highly reactive to LSAs in general. Manitoba is a 'low-low' province. Another contrasting example is New Brunswick which has high negative reactivity effects, but also high positive reactivity effects. It is a 'high-high' province. Ontario is a low-high, and in the least desirable way: low positive reactivity, and high negative reactivity. In some cases, though, test-taking strategies may be necessary part of instruction:

For the grade 9 math EQAO the results went up significantly primarily in students with an IEP (individualized education plan) because the other resource teacher and myself had spent a lot of time doing direct instruction on how to do well on the test. And that may be coping strategies if you are tired, it may be making sure your calculator works, or you know how to work your

calculator. Things like that as opposed to teaching to the test, some of the other skills students need.

- ON, High school English consultant, female

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

I will be really honest with you. I have been kind of responsible for overseeing literacy [test] preparation. . . I treated it as a very custodial exercise. I mean, I looked at the test and I sort of, umm, I put together lesson plans that I thought any teacher would be able to deliver to their students. . . And we still do this, I mean, even though we started embedding it more in what we do, every year we have done exactly the same thing. . . We do all of our training in grade 10 classes period 1, because those are the kids you have to hit and I want teachers to have ownership. So what happens is we teach to the test. We basically have, I have seven lessons that are very narrow. . . They do one a week for seven weeks leading up to the test.

- ON, High school principal, male

So while Ontario appears somewhat off-balance with regard to the type of reactive strategies reported by the sample of teachers in this survey, the stated mission of EQAO is to 'build capacity for appropriate data use.' By the measurements employed in this study, that is not being done. Negative reactivity effects (strong in Ontario) are the least appropriate uses, and positive reactivity effects (weak in Ontario) are the most appropriate. They are in the wrong proportions for one of EQAO's main goals to be achieved. It has been noted that only one Canadian province has more positive reactivity effects than negative, and this is troubling. However, Ontario remains unique in that it has far more negative reactivity effects than those that are positive. If positive reactivity is the high road to academic success and learning that transcends the walls of public schools, then there needs to be a shift in implementation practices in Ontario.

Prince Edward Island

- DV1 - positive reactivity

PEI respondents are close to national averages for positive reactivity with slightly more 'strong positive' than is the norm (8% more) and less neutral positive (8% less). The average score is 2.82, higher than the 2.50 national average.

- DV2 - negative reactivity

For the negative reactivity metric, PEI teachers show more negative effects, but most of these are in the moderate negative grouping (62%), and a lower proportion of strong negative reactivity (32%) than is true nationally. The average score in PEI is -2.90, slightly higher than the national average of -3.04.

- DV3 - total reactivity

Overall, PEI is about average in total reactivity with 50% of respondents on the moderate/neutral side of the fence, and the other half on the strong/very strong side. The average rating in PEI is 5.72 while 5.53 is the national average.

- DV4 - net reactivity

Net effects show an even starker picture of PEI's reactivity practices. The balance tips slightly negative (18% moderate negative as compared to 12% moderate positive), but 71% of respondents fall into the neutral group after cancelling out happens. This is the second highest provincial rating for net neutral effects after Saskatchewan. The average balances to -0.07, just into negative territory, and higher than the national average of -0.54.

- Summary of reactivity effects

Like most provinces, negative reactivity effects predominate in the PEI sample from this study. The amount of negative reactivity here is not extreme but it does outweigh the positive effects. A related concern is that reactivity in general is not strong in PEI schools. Teachers are not certain about taking large steps to positive or negative reactive strategies. This indicates that the ministry (and/or the board, and/or the schools) has not made explicit enough any expectation to use the data in appropriate ways. While directing resources and identifying students needing interventions are part of their mandate, teachers apparently are not prepared for this task or have other priorities:

We struggle with finding a balance. You want them [students] to do enough so that they are prepared, but you don't want to teach to the test. . . So how do you find that balance when you only have so many hours in a day?

- PEI, Elementary homeroom teacher, female

Like I said . . . we do not use it [the data] to frame the way we teach, and we question and have questioned in the past why we are even

doing them. Like, what is the benefit from that? You know, we struggle with that. - **PEI, K-9 school principal, male**

Accountability in schools is a, it's a political realm . . . you are dealing in union, you're dealing in public perception, you're dealing in politics, you're dealing in re-election, and no. I think when you, teachers, are professionals and when you work professionally with teachers there's many ways to assess students. . . It's just a very convenient method whereby spin-doctors and people can, you know, bureaucrats in rooms, no offence to them, bureaucrats in rooms can look at a number and they don't see a face, and they see a trend and think that they can address an issue. It is all bigger than that, you know. I see a face, and it has a home, or a lack of a home. You know, there are many other issues that are involved in teaching. As a teacher first and an administrator second, no . . . I say that I don't need a provincial assessment [for accountability].

- **PEI, K-9 school principal, female**

Prince Edward Island's program seems to be, as compared to the other provinces, one that is specifically focused on providing teachers with information they can use to direct resources and interventions. This conclusion is based on that fact that provincial assessments in PEI do not occur at the high school level (grades 10 – 12) and there are no graduation requirement tests written at any level. For the sake of comparison: Ontario has the OSSLT (Ontario Secondary School Literacy Test) and New Brunswick has the ELPA (English language Proficiency Assessment) which must be passed to graduate; Québec, Alberta, British Columbia, Manitoba and Newfoundland and Labrador schools have exit exams (written in grades 10, 11 or 12) that make up a large portion of a student's final mark; Nova Scotia has similar set of new mandatory assessments for core subjects for the 2013-2014 school year; and Saskatchewan is in the middle of re-inventing provincial assessment, so has nothing in high school beyond an affective questionnaire (this will likely change) except in cases where teachers are not accredited and departmental exams must be written. Prince Edward Island alone focuses on younger students and uses no exit exams of any kind. The policy choice to create useable data for improving student learning (which I must assume was a conscious choice) falls down at the implementation stage, though, if teachers are not ready, willing, and able to use the data in constructive ways. Rather than writing new assessments for more grades, this fine tuning and clarification of the existing structure should be the focus of the assessment branch moving forward.

Grading schools, I was working in a local high school and I know that they had different ratings. . . Well, we had no idea was the criteria was [sic] for that. And, you know, one school had a 'D' mark one year and a 'A-' the next year, but we had no idea. . . like you didn't have a sudden culture shock change.

- PEI, High school Math teacher, male

I don't think we are preparing anything specifically for the test. I think we are just following along with what is expected to be covered in the curriculum documents. . . Not a lot preparing them specifically to write the test, I would say no.

- PEI, K-9 school principal, male

An impetus to work with data for instructional change in different ways was evident here, as in all jurisdictions, but survey data do not support the idea that the methods were agreed upon or common to most teachers.

We would identify areas we wanted to show growth or improvement. So in our school we would have had the kids tested in grades 3, 6 and 9 in mathematics. So it might be because we are a small school you can boil it down to. . . students A, B, and C are struggling in a particular area in the grade 6 assessment. You know, what are some of the things we can do to try to help them in grade 7 and then prepare them potentially for like the grade 9 [assessment]? - PEI, High school Math teacher, male

We do give practice sessions and kind of familiarize them with bubble answers and how to fill in, you know, the multiple choice answer and that type of thing.

- PEI, K-9 school vice principal, female

It seems clear from survey and interview results that some direction is needed and in many cases requested.

Québec

- DV1 - positive reactivity

Québec has the third highest provincial rating for positive reactivity. Moderate effects reach 45% of respondents while strong positive effects were reported by 38%. The average score for this metric was fourth highest (2.78) better than the national average (2.50).

- DV2 - negative reactivity

Negative reactivity is also found in Québec, and strong negative effects outpace the national average by 11%. Moderate negative effects (38%) are less than the national average by 5%. The average is -3.52, which is the second lowest provincial average and lower than the -3.04 national average.

- DV3 - total reactivity

With high numbers for strong positive and strong negative reactivity, Québec is the second most reactive province looking at total effects. Strong reactivity describes 46% of respondents and very strong reactivity describes 26%. The average score is 6.29, just behind Alberta's 6.39 and higher than the national average of 5.53.

- DV4 - net reactivity

Net reactivity paints an odd picture in Québec. There are no strong negative effects nor moderate positive effects in evidence. Yet 3% of respondents fall into the strong positive category, 59% in net neutral, and 38% in the moderate negative grouping. The balance thus tips toward the negative side and the average overall score is -0.74, lower than the national average of -0.54.

- Summary of reactivity effects

As the province with the second highest ratings for total reactivity, there is clearly strong support for the use of positive reactive strategies and even stronger support for those strategies classified as negative reactivity. The LSA data are absolutely being used in Québec to alter instructional practices. Having discussed the nature of the Québec assessments with teachers in that province, this was not a surprising finding: the data are meant to change instruction.

You have to make sure that you have covered most everything. I would say the things that comes back the most are fractions, geometry. . . basic facts. But the way it is all worded it is very professional. The kids often have to read it at least five times before they even start.

- QC, Middle years homeroom teacher, female

So we just program as best we can and we do make sure, and it does change what we do. There is not a teacher in Québec, I don't think, who would teach a formal reading response to grade 11 students other than the fact that that is 1/3 of their global mark at the end of the year and they have to do that particular, quite artificial task. We spend an inordinate amount of time on a particular task which is not something we would do on a formal

basis, I don't think. It takes away from a lot of the other things we would like to do in the cycle [high school].

- **QC, High school English teacher, male**

It is also clear from interviews respondents that exactly how they are supposed to use the data is only somewhat clear at the best of times and an 'unknown unknown' at the worst. This researcher found that online inquiries to discover information about assessment in Québec (from the English language versions of provincial websites) were frustrated at every turn. One respondent from Québec related that teachers, administrators and parents were all unaware of the goings on behind closed doors at Québec's assessment branch (related to marking and reporting scores). If accountability is one rationale for the provincial assessment policy (and I cannot unfortunately confirm or deny this supposition), then the ministry might consider their own practices as a starting point for reform.

I'm going to try and say this very carefully. We care very much about how our students do; we very much want our students to succeed. I think we feel that the provincial exams and how they are dealt with are almost completely out of our control. . . So we teach what we know the students need to know, try to have in as much of the other things that we think they should know to be well-rounded people and then we kind of let the chips fall where they may when the exam comes around.

- **QC, High school English teacher, male**

So all of that [how marks are calculated] is opaque - from the level of the ministry of education there is no accountability, no transparency. . . - **QC, High school English teacher, male**

Saskatchewan

- DV1 - positive reactivity

Saskatchewan has a lower average score in this metric (2.17) than is true nationally (2.50). There is somewhat more 'neutral positive' reactivity in evidence than the national figure (32% compared to 26%) and less strong positive by a 7% margin.

- DV2 - negative reactivity

Saskatchewan has the second lowest provincial proportion of negative reactivity and the second highest average score in this category (-2.38 compared to the national -3.04). The neutral negative effects here are correspondingly higher than the national average (by 16%) and the strong negative is lower (by 22%).

- DV3 - total reactivity

The low numbers for positive and negative reactive effects make Saskatchewan the second least reactive province as 66% of respondents fall into either the neutral or the moderate groupings, where the figure is 42% for the national data. The average score is 4.55, while the national average is 5.53.

- DV4 - net reactivity

Saskatchewan has the largest proportion of net neutral reactivity at 78%. The balance of positive and negative tips slightly to the negative, therefore, based on the 14% of moderate negative and 2% strong negative opposed to only 6% moderate positive. The average score is -0.21, where the national average is lower at -0.54.

- Summary of reactivity effects

Saskatchewan may be the second least reactive province, but with perhaps the best reason: they do not currently have a battery of provincial assessments field tested and implemented in schools. There are pilots of assessments for Pre-K through grade 3, and these are the only LSAs that could conceivably be fully implemented over the next couple of years. With no middle years or high school assessments in development or piloting (at least that are known to the researcher), these higher level assessments, if they are to return, are 4 to 5 years off.

The early years assessments being piloted in 5 school divisions now promise to be very different from the Assessment for Learning (AfL) tests that came before in terms of both design and target grade level. The researcher cannot speak authoritatively about these new tests, but they rely more heavily on technology than other jurisdictions, and have a 'holistic' approach that does not correspond with the LSAs currently implemented elsewhere in Canada. A fair evaluation of their effectiveness may be 3 to 4 years away, as well.

So the reactivity effects noted in survey responses here were based on the older, now obsolete AfL tests, and has no bearing on the assessments that will be coming in the future. If these results say anything, it is that the AfL model was not effective at promoting change in instructional methods and that the ministry is likely on the right track to replace them with something that provides more actionable data and the means to do that. The question was posed whether provincial testing improves teaching and respondents answered:

No. I think collaborating with other teachers improves teaching often. And I think you have to be willing to realize that there is change that can be good out of changing and to improve on your teaching practices, but I don't think that provincial testing does [that]. – SK, Middle years homeroom teacher, female

No, not [testing] at the provincial level. I'm all about formative assessment to be a better teacher. . . I want my teachers to know their kids and to know what they need and to teach them at that level. But I don't want restrictions on that. . . . I think it is important for teachers to know where their kids are and what they know so we can be better teachers and the kids can be better learners.

- **SK, Elementary school principal, male (a)**

The administrators that did a good job of connecting the assessments that we had to fulfill ministry requirements, connecting those to like my own professional goals and plans and professional development. Then I agree[d] with the whole process. But then I have also taught under administrators who did the exact opposite, where it was just, we had to do it. It doesn't matter, the results. It doesn't matter how the kids do. It really means nothing, it is just checking something off on a list that we have to do. Then I don't agree with it. . . I think it really depends on how the school is using it.

- **SK, Elementary school principal, male (b)**

The pilot tests are theoretically 'linked to activities' and provide the means for 'immediate action' from not just teachers, but also parents. This is a promising sales pitch, and it will be most interesting to see how they fare in classrooms.

3.5 Conclusions

Teacher reactivity in Canadian schools is a key driver in instructional change and improvement. There are no comparable effects noticeable in international (PISA) or even national (PCAP) tests since these seem to have no stakes involved (resulting sanctions from school or division level bodies are almost unheard of in Canada), no incentives for improvement aside from pride, and there is no general consensus on how improvement on these metrics might be achieved considering the sample-style of the test and the lack of preparation materials.⁷² PISA does not appear to affect day-to-day teaching or learning.

⁷² Michael Corbett (2006) is cited in Morgan (2009) as saying PISA is "divorced from the reality of the classroom and the context in which schools do their work." Breakspear (2012) indicates the PISA scores influence *national* accountability, not classrooms. Uljens (2007) relates the idea that PISA is 'soft power' and participating states self-impose changes at the national level.

If you want large-scale assessment to have an impact on schools, then you should disseminate results and discuss them with teachers, parents, school boards, and so forth on the local level. . . Even if all are well informed, the question is, what do you do with the PISA results? How should our restructuring of schools take place so that better results can be achieved? (Ritzen, 2013, p.23)

What is true is that these international gauges of educational quality do matter to the provinces: all provinces make reference to PISA and or PCAP when spelling out their own assessment policies. These tests seem to be very effective in promoting a culture of testing in schools as part of a more general 'audit culture' (as it has been called) in government at large.⁷³

Yet serious doubts about testing policies arise from the results of this survey. Some provinces have little success in having teachers use the results in any way. Nova Scotia, Saskatchewan and British Columbia are the least reactive provinces. Some provinces do not have a problem with total reactivity scores, but instead see very prominent negative reactive effects. In provinces that give exit exams or minimum competency exams in high school, strong negative effects are most strongly pronounced (listed from the highest proportion of effects: Ontario, Québec, Newfoundland and Labrador, and Alberta). In all provinces (even including the barely net positive Nova Scotia) the problem is that teachers do not distinguish very readily between positive and negative reactive practices. In nine Canadian provinces there are more negative reactive strategies in use than positive ones.

If there is a critique of the conclusions drawn from the data in this chapter it would most likely be that the additive scale for positive reactivity has a low value for Cronbach's alpha. Used as an index of reliability, the score of 0.62 is less than the 0.7 considered the threshold for reliability. The value for negative reactivity was above this threshold, and despite the alpha score, the author has confidence that the results do, in fact, accurately describe the reactive practices of teachers, both positive and negative. Further research could be done to develop scale with a higher alpha score.

The researcher was asked early on during the development of this project whether teachers would really be honest about using negative reactivity practices – 'wouldn't they just answer the way they were expected to?' Ashton, Buhr and

⁷³ Shore, 2008. p.287-288: "The rise of audit culture in Britain can be traced to the reforms of the 1979 Thatcher government, which sought to reduce public spending, 'roll back the state', and increase the efficiency of public servants by subjecting them to the simulated disciplines of the market. Ministers insisted that the private sector was regulated effectively by market mechanisms and that this model should therefore be applied to the public sector."

Crocker (1984), for example, found that teachers' sense of efficacy was norm-referenced – they would not admit to socially undesirable actions. So these results betray either a great amount of honesty, or what is more likely, some legitimate confusion about those educational practices that are both ethical **and** provide the best educational outcomes for students. This perspective was examined by Popham (2001) who stated that teachers rarely considered the appropriateness of their test preparation activities. One can always debate the relative merits of specific instructional methods, but these choices are not always easy to make in one's own classroom. In broad strokes, though, there is consensus that the following strategies do not best serve learners: (a) narrowing the curriculum; (b) teaching skills that are useful only in exam situations; (c) restricting the use of varied and creative assessment methods; (d) side-lining non-core subjects; and (e) reviewing previous tests seeking an easy way to prepare students without deeper understanding of the content. These 'teaching to the test' practices are today used more commonly in nine out of ten Canadian provinces than positive reactive strategies.

The chapters that follow will take up the trail of reactivity effects to see what combinations of policy factors seem to lead to more total reactivity, promote more positive reactivity, and create the conditions where negative reactivity is employed.

3.6 Charts and tables

Figure 3.4: Positive reactivity scores, as self-reported on surveys, given numeric values and added together give each respondent a score from 0 to 5. Positive reactivity practices were rated as used a great deal (1), somewhat (0.5), or not at all (0).

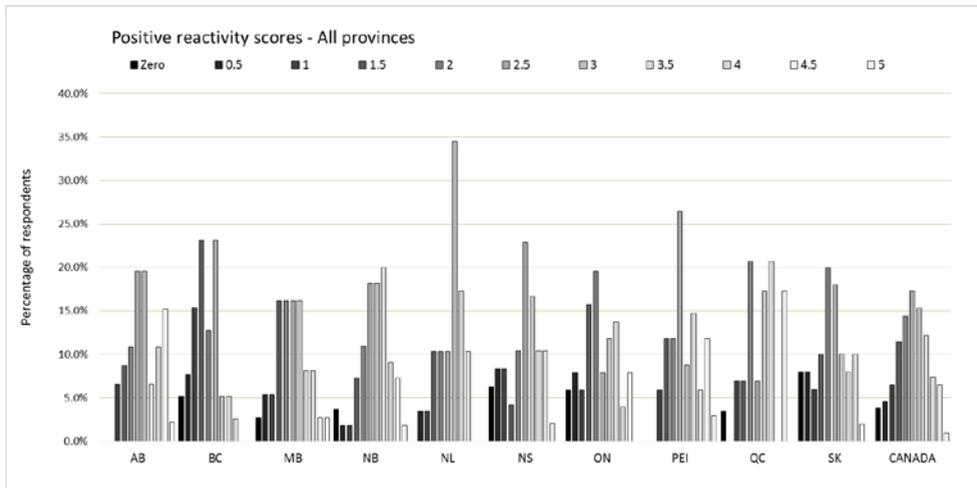


Table 3.5: Distribution analysis for national data from figure 3.4. The national distribution, while under-dispersed ($D = 0.5463$), this is possibly a result of a small range of values – the range is only 0 to 5. There is no significant skewness (-.1468) or kurtosis (2.4764).

National data for positive reactivity effects		Observations: 418
Mean: 2.4952	Standard deviation: 1.1676	Variance: 1.363
Range: 0 to 5	Skewness: -0.1468	Kurtosis: 2.476

Figure 3.6: Aggregating the data from figure 3.4 into three groupings (neutral positive, scores 0 to 1.5; moderate positive, scores 2 to 3; strong positive, scores 3.5 to 5).

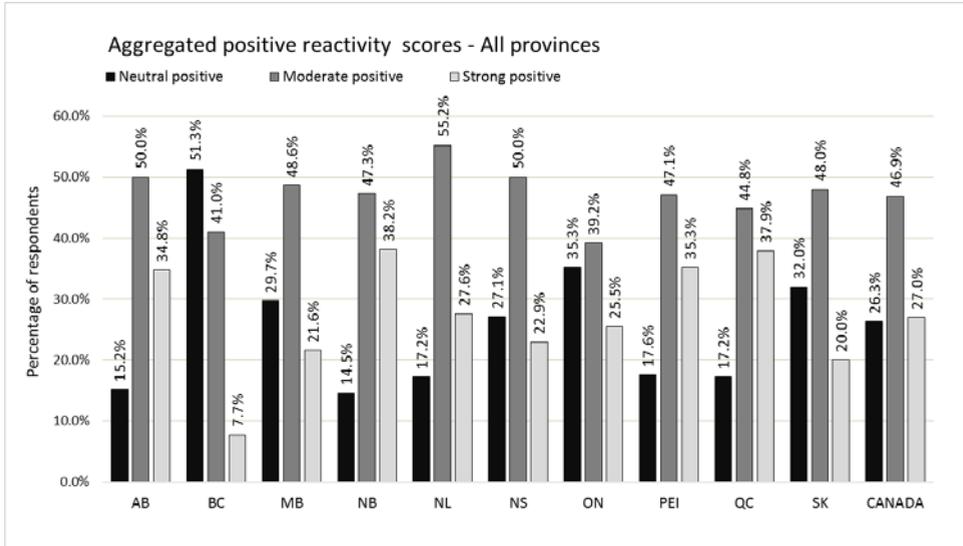


Figure 3.7: Negative reactivity practices, as self-reported on surveys, were given numeric values and added together give each respondent a score from 0 to -5. Negative reactivity practices were rated as used a great deal (-1), somewhat (-0.5), or not at all (0).

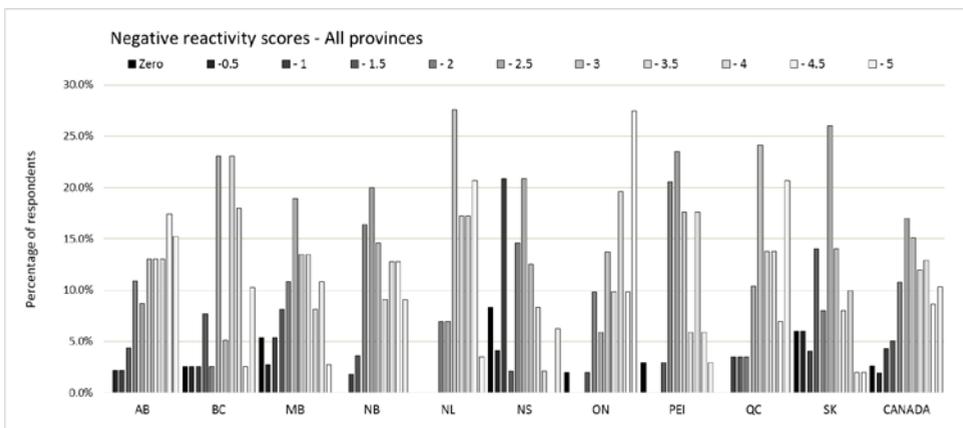


Table 3.8: Distribution analysis for national data from figure 3.7. The national distribution has a higher mean score than positive reactivity effects (absolute values: 3.036 and 2.495 respectively) and is skewed left (.3181), indicating high proportions of negative reactivity effects. It is also under-dispersed ($D = -0.5201$) since the range of possible values is quite small (0 to -5).

National data for negative reactivity effects		Observations: 418
Mean: -3.0358	Standard deviation: 1.2566	Variance: 1.579
Range: 0 to -5	Skewness: 0.3181	Kurtosis: 2.592

Figure 3.9: Aggregating the data from figure 3.6 into three groupings (neutral negative, scores 0 to -1.5; moderate negative, scores -2 to -3; strong negative, scores -3.5 to -5).

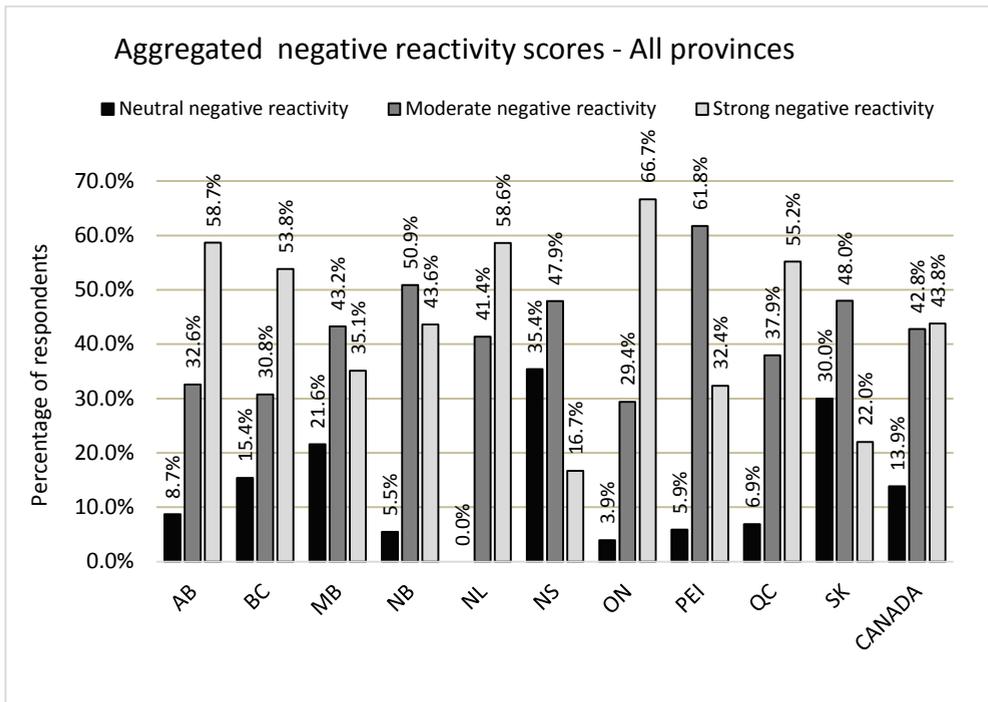


Figure 3.10: The added absolute values of positive (0 to 5) and negative (0 to -5) reactivity scores are plotted above for respondents from all provinces. Note that all total scores above 5 indicate the use of both positive and negative reactivity practices.

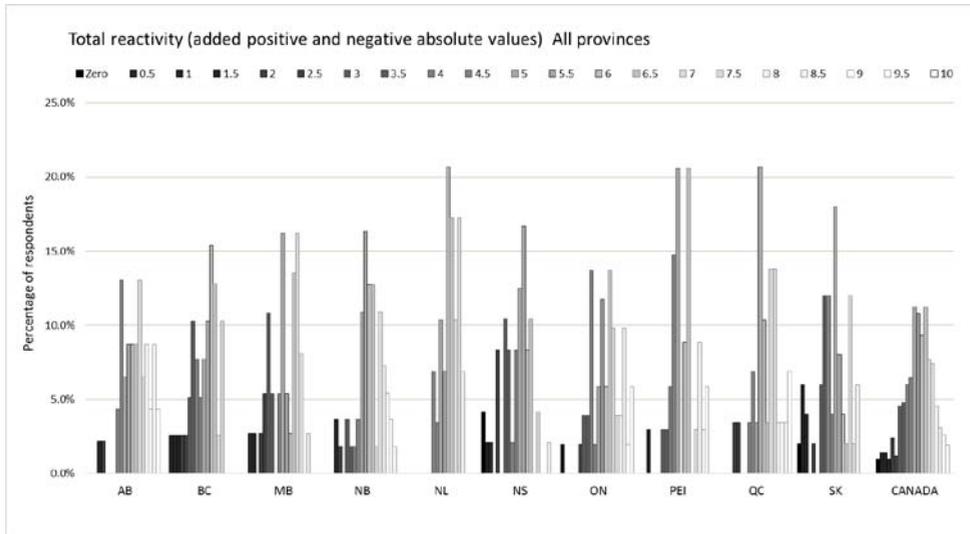


Table 3.11: Distribution analysis for national data in figure 3.10. The national distribution skews left (-0.4417) of the mean, indicating high levels of total reactivity effects and one small tail on the left denoting fewer neutral effects. With a mean score above 5 (5.5311), the presence of simultaneous positive and negative reactivity effects is shown to be quite common. The variance ratio is better in this measure with the full range of reactivity response values included ($D = 0.7234$), but remains slightly under-dispersed. What seems clear is that while neutral reactivity is present in almost all provinces and nationally, the levels of simultaneously high positive and high negative reactivity show that most teachers do not or cannot distinguish between the relative merits of these two sets of practices.

National data for total reactivity effects		Observations: 418
Mean: 5.5311	Standard deviation: 2.0004	Variance: 4.001
Range: 0 to 10	Skewness: -0.4417	Kurtosis: 3.062

Figure 3.12: Aggregating the data from figure 3.10 into four groupings of reactivity effects. A score 3 or lower indicates neutral reactivity (double the cut-off value for neutral positive and neutral negative scores). Moderate reactivity encompasses scores 3.5 - 5, strong reactivity covers scores 5.5 - 7, and very strong reactivity includes scores 7.5 and higher.

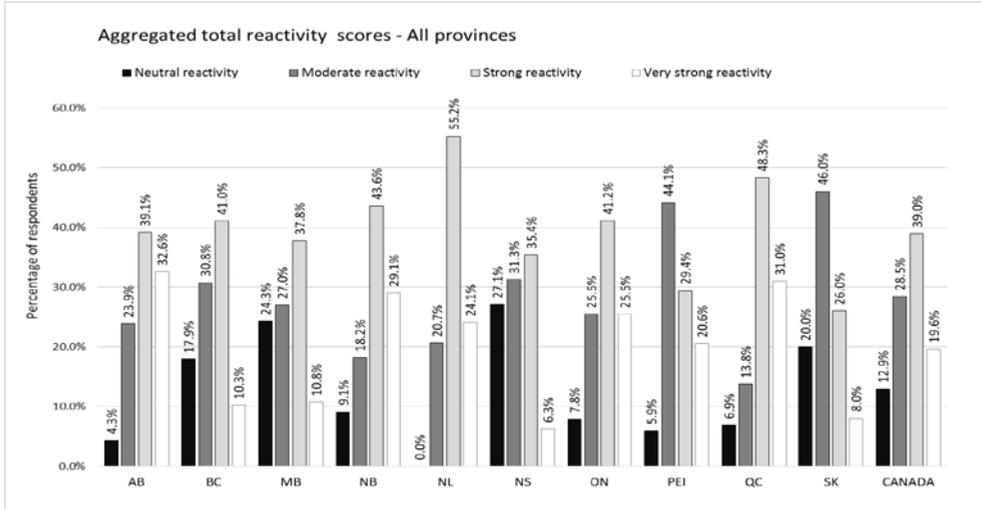


Figure 3.13: Net reactivity effects add the positive score and the negative scores meaning that the positive and negative values might cancel each other out. These are aggregated data on net reactivity, with groups from 0 to ± 1 (neutral), ± 1.5 to ± 3 (moderate), and ± 3.5 to ± 5 (strong). This chart shows these aggregated data for all provinces and nationally.

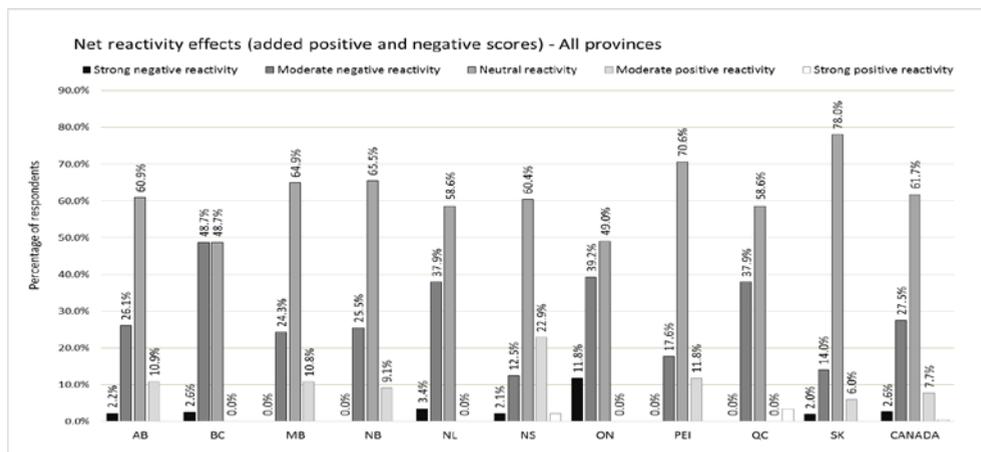


Figure 3.14: The graph shows detailed national data for net reactivity (explained above).

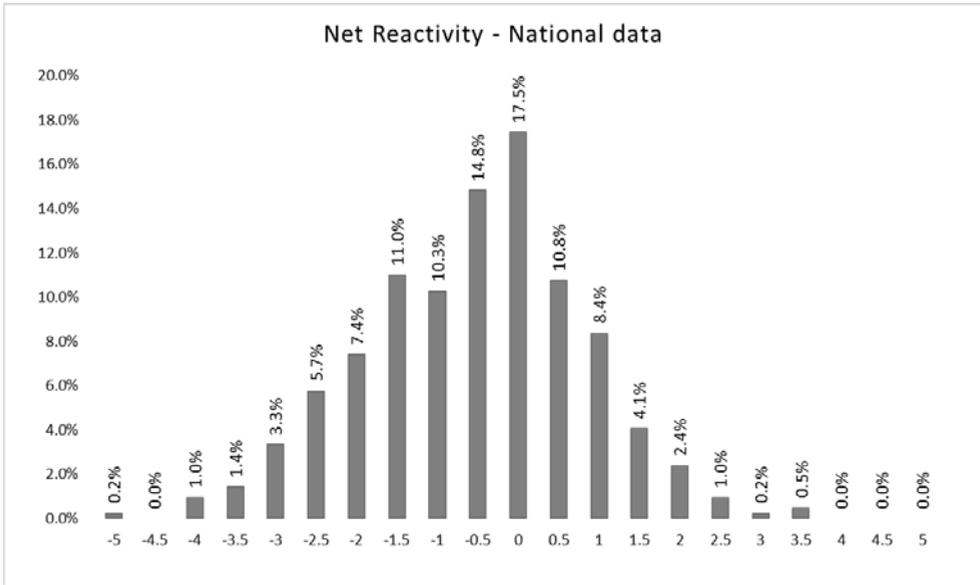


Table 3.15: Distribution analysis for national data from figure 3.14. This national distribution of net reactivity effects is fairly normal, but skewed slightly right (-.1414) and a normal kurtosis (3.0335). The mean is slightly into negative territory (-0.5407) with a standard deviation of 1.3723. The mathematical cancelling of positive and negative reactivity scores means the tails (extreme positive only or negative only scores) are quite narrow.

National data for net reactivity effects		Observations: 418
Mean: -0.5407	Standard deviation: 1.3723	Variance: 1.883
Range: -5 to 5	Skewness: -0.1414	Kurtosis: 3.034

Test design, results data, and attitudes about testing

4.1 Introduction

Large-scale assessment is a massive, costly, and complex undertaking. It is common to all Canadian provinces and most industrialized nations, yet the very complexity of large scale testing means that *how* it is done (the subjects tested; the grade levels chosen; the test instruments; data analysis; and the return of data to stakeholders) is quite different across different jurisdictions. These policy choices will be examined in light of their effects on teachers use of LSA results data to improve instruction (reactivity).

This chapter is laid out in the following way: (a) a literature review discussing previous work on both the policy-level and the classroom-level high-level policy choices made about tests and the LSA results data; (b) a discussion of the findings from the researcher's survey of teachers all across Canada; and (c) conclusions are presented. These chapter-specific results will address several important independent variables (IVs): IV1 – time of data return; IV2 – aggregated / disaggregated data; IV3 - item types; IV4 – presentation of results; IV5 - data clarity; IV6 - ability to act on data; IV7 – results use for school accountability; IV8 - results use for student accountability; IV9 - results use for school improvement; IV10 - negative attitudes about testing; and IV11 - appropriate uses for the data.

4.2 Literature review

4.2.1 Policy-level considerations

International standardized testing is more prominent now than ever before and 2015 will see a new round of PISA tests with the almost certain resultant political firestorms. PISA creates controversy partly because no country (or province) wants to see their results decline, and partly because the OECD allows participant nations to set policy goals to 'correct' their faults:

"As PISA is mainly focused on ranking of participating countries and not very interested in explaining differences between them, the burden of producing explanations is left to the participating nations, their governments, educational administration and the media."
(Uljens, 2007. Ppg. 12)

In the author's home province of Saskatchewan PISA results were named as a reason for concern since it put us "at a serious disadvantage" compared to better-

performing provinces and nations.⁷⁴ The policy movement of countries that do participate, despite the 'soft power' employed by PISA, seems to be all in one direction – toward more testing (Uljens, 2007; Morgan, 2009). Some current research goes yet further to explain that Canadian provinces see three common effects from international testing and the wide-spread publication of results:

Three patterns were particularly notable: (a) the salience of test scores is in part tied to relative performance across provinces; (b) curriculum reforms often intensify for tested subjects in response to international test results; and (c) school renewal efforts tied to international test results are heavily influenced by geopolitical forces. (Volante, 2013. pg. 173)

So international testing has its role in the promotion of more national and provincial testing. Canadian provinces have all taken up the practice despite the objections common among educators and researchers alike that the tests have unintended consequences for schools and de-professionalize teachers (Amrein, Berliner & Rideau, 2010; Runté, 1998). While it is not commonly disputed that schools as public institutions should be transparent to some degree, what is questioned is the level of accountability provided when the tool used is a standardized test (Morris, 2011). There is an important distinction to be made about the two faces of accountability: the public/political and the professional (Møller, 2008; Ben Jaafar & Anderson, 2007; Darling-Hammond, 2004). Accountability testing is most commonly ascribed to the first camp, as a politically motivated process, whereas professional accountability refers to teachers living up to the standards of their profession and being responsible, reflective educators (Hargreaves et al., 2009). Political accountability also tends towards the use of political mechanisms to achieve policy goals:

The focus on accountability use standards, assessment, rewards and punishment as its core drivers. It assumes that educators will respond to these prods by putting in the effort to make the necessary changes. It assumes that educators have the capacity or will be motivated to develop the skills and competencies to get better results. (Fullan, 2011. pg. 8)

It is not clear that the political accountability model has proven effective at improving teaching.

⁷⁴ Saskatchewan Ministry of Education, 2012b

Another prime consideration at the policy level is the test instrument itself. When multiple and often conflicting goals are codified in a single policy it becomes quite complicated for even the cleverest psychometricians to devise the one instrument to meet them (Hamilton, 2003; Klinger, DeLuca & Miller, 2008). Using the data from provincial websites, the researcher identified nine different purposes/aims explicitly stated or implied across education ministries (see **Figure 1.2**). These included: (a) reporting and accountability; (b) graduation requirements; (c) improving central and/or local data-based decision making; (e) interventions for struggling students; (f) improving assessment literacy or professional development for teachers; (g) a response to PISA or PCAP scores; and (h) the monitoring and improvement of achievement results. The fewest of these nine goals expected from any province was five (in PEI) while Ontario expected all nine to be met to some degree. It was noted in chapter one that "tests used primarily for curriculum advancement will look very different from those used for accountability" (Anderson cited in Mehrens, 1998. pg. 8). It seems that the message about multiple purpose tests has not been well-received in most Canadian provinces.

Two other key considerations in designing large-scale assessments are validity (if the assessment covers stated outcomes and content) and reliability (if assessment tools provide consistent data over repeated occasions). Validity in testing is examined in great detail in the work of Koretz (2002, 2005, 2009, and 2010) where forms of reactivity such as coaching and score inflation are examined and documented. Validity is not always a primary concern. It might be seen as a case of the putting cart before the horse, but many tests are designed (these metrics of accountability) based on what is most efficient, most cost effective, and provides reliable, comparable data.

Clearly, external tests should closely mirror the state, provincial, or territorial curricula and include higher-order thinking tasks that are essential within our knowledge economy. Ironically, when broad curriculum coverage and critical thinking skills figure more prominently within external tests, teaching to the test actually becomes a desirable objective. It is quality, not ease of development, administration, and scoring, that should always be the first consideration in the selection of any large-scale assessment measure (Volante & Cherubini, 2007, p.5).

Sample style assessments, where a chosen sample of students or schools takes part, are much cheaper and less time intensive than a census style test. But these are most often used only for system diagnostic purposes since they by design provide no individualized data (Morris, 2011; Nagy, 2000). The relatively low-stakes of sample style tests also makes them less relevant to school level staff

(Cimbricz, 2002). This type of test provides results at least as valid as high stakes census style tests. Koretz notes that there are many upsides to improving census-type testing practices, but also increased costs:

One essential step for the measurement community is to adapt principles of test design to reflect the risks posed by high stakes testing. This will require limiting unnecessary, predictable recurrences and omissions. While some have suggested richer, less predictable sampling of content and skills, this would be necessary but insufficient. It will also be necessary to limit predictability of non-substantive performance elements in order to lessen coaching. Decreasing predictability will improve the incentives created for educators and students. It will also make inappropriate coaching and reallocation more difficult and thereby lessen the risk of degraded instruction and score inflation. However, this change in design, while straightforward in principle, may prove difficult in practice. It will obviously increase test-development costs. (Koretz, 2009, p.13)

Canadian LSAs are generally low stakes for teachers and students alike (Volante, 2006). Having said this, there are certainly implications that follow results being released. With the purpose of improved classroom instruction and “strengthening school capacity” in mind (Saskatchewan Ministry of Education, 2007), there are many documented instances of the opposite occurring: (a) teaching to the test and curriculum narrowing (Morris, 2011); (b) less content depth in classrooms and little of the resulting information being used to improve classroom practices (Volante, 2006); (c) compromising standards of good teaching to meet accountability goals (Mintrop et al., 2009); (d) minimal substantive insight to teachers provided from results (Shepard et al., 2011); and (d) 'borrowing' time from non-core subjects to cover assessed material (Ungerleider, 2003). None of these practices is a proper pedagogical strategy to improve learning, nor do they have any lasting impact on this learning (Volante, 2004). An alternative is to rely on multiple measures for assessment purposes and to balance the weight given to provincial tests which cannot serve all the masters to which they have been assigned. Reliance on teachers' own classroom-generated marks which may better reflect the curriculum and are not averse to in-depth learning or performance assessment may be one way forward (Koretz, 2005).

Test development and implementation are obviously complex functions that Canadian provincial ministries are tasked with completing to the satisfaction of parents, students, teachers, administrators and superintendents. Getting the right mix of 'actionable' data without extreme cost in time and money is their goal:

Schools are typically awash in many different types of assessment data, and it is a significant design challenge to constrain the various assessments to produce a reliable and shared measure of student learning. School learning assessments exist at many different levels and serve different purposes. Schools and districts are held accountable by governments for documenting student learning in terms of summative standardized tests. Local schools and classrooms receive and design a wide variety of formative assessments, ranging from benchmark assessment systems to teacher developed quizzes and homework checks, to monitor the learning process. A challenge of formative feedback system design is to establish a direct link between interventions and assessments to create actionable information for faculty and staff. (Halverson, 2010, p.140)

4.2.2 Classroom-level considerations

Cizek (2001) notes that there are many serious negative consequences of external testing: (a) the reduction of time for regular and thorough instruction; (b) the neglect of domains not covered in LSAs; (c) using instructional strategies rooted in test designs; (d) limiting broad-based and productive instructional opportunities; (e) poor morale noted in school staff; and (f) the use of LSAs as a form of punishment for students. Some of these topics are extensively covered in the literature and are frequently cited as 'unintended consequences' and: (a) teaching to the test; (b) coaching; (c) curriculum narrowing; and (d) outright cheating (Volante & Cherubini, 2007; Luna & Turner, 2001; Simner, 2000; Fehr, 2008; Volante, 2007; Darling-Hammond & Rustique-Forrester, 2005).

Between clearly unethical assessment practices and those that are beyond reproach is a gray area where teachers have a hard time distinguishing if what they are doing violates their professional ethics. Koretz and Jennings (2010, p.18) note that often teachers have a difficult time with this judgement:

The dividing line between appropriate and inappropriate preparation is not always entirely clear in a specific case, but the general principle is unambiguous. 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.

The inability to make this judgment is no small issue especially when trained professionals cannot themselves determine the appropriateness of test preparation activities. Therefore the school system is at risk of both overt and covert score manipulations. Koretz (2005) notes that reallocation of resources, a seemingly ethics-free choice, may also include practices like curriculum narrowing and increased focus on the tested domain.

This gray area between ethical and unethical is perhaps rooted in the assumption that accountability assessments should not have any negative consequences but this view has been shown to be naïve (Mehrens, 1998). The fact that divisions often support 'gray area practices' goes even further to help us understand the problematic side of test reactivity.⁷⁵ **Figure 4.1** shows an adaptation (it has been edited and supplemented) of a table from Haladyna, Bobbit Nolen & Haas (1991, p.4) which breaks down many of the positive and negative reactivity elements cited above in terms of ethics and their effects on student outcomes.

Figure 4.1: Ethical and unethical educational practices and their effect on the number and variety of student outcomes

Test preparation activity	Degree of ethicality	Effect on student outcomes
Training in test-taking skills	Ethical	Narrowing
Checking answer sheets to make sure that each has been properly completed	Ethical	None
Increasing student motivation to perform on the tests through appeals to parents, students and teachers	Ethical	None
Developing a curriculum based on the content of the tests	Unethical	Narrowing
Preparing objectives based on items on the test and teaching accordingly	Unethical	Narrowing
Presenting items similar to those on the test	Unethical	Narrowing
Dismissing low-achieving students on testing day to artificially boost test scores	Highly Unethical	None
Presenting items verbatim from the test to be given	Highly Unethical	Narrowing

⁷⁵ Koretz (2009, p.9) notes: "Indeed, state and district education agencies often encourage it [reallocation within subjects], either implicitly, for example, by providing previous test forms as a guide for study, or explicitly, by informing educators which portions of the domain are given substantial weight by the test. This latter approach often goes by the disarming name of 'power standards.' "

This helps illustrate the mixed bag of instructional practices that are employed in the name of, and in reaction to, high stakes large-scale student assessment.

Not all reactivity is negative, though. Positive effects of reactivity have been noted by several scholars and studies. If test results are used well, they can be a guide to instruction, and they can ensure that the entire curriculum is covered in depth. This type of broad curriculum coverage is regarded by Volante (2004) as a better preparation for future success. This is an example of reflective practice. Where LSA data is used as a means of self-assessing what has been done in classrooms and adjusting for improvement, it is an important element of personal professional improvement (Loughran, 2002). These data can guide instruction in positive ways.

Monitoring, questioning, and providing feedback are means by which teachers can stimulate achievement gains on LSAs and in classrooms (Gibson & Dembo, 1984). Nagy (2000) argues that teachers, as responsible professionals, will act on appropriate data without any coercion whatsoever because it is just good practice and professional competence to do so. Volante (2005) also notes the many ways individual teachers can use LSA data: (a) to identify strengths and weaknesses in the curriculum; (b) to assess their own pedagogy; (c) providing a chance to seek out relevant PD; and (d) giving evidence to judge quality of school programs and the allocation of resources.

Positive reactivity from LSA regimes can also result in improved teamwork and collaboration between teachers. Cited in Cimbricz, Grant (2000) notes that:

Several teachers, especially elementary and high school math and English teachers, for example, cited greater collaboration with their peers. The development of informal networks and relationships, therefore, was reported as one of the key benefits stemming from the changes made in the state-mandated testing program. (Cimbricz, 2002, p.12)

Darling-Hammond and Rustique-Forrester (2005) 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; (c) to provide incentives for administration to offer opportunities and resources for PD; and (d) to indicate what domains are most important to teach and learn by setting clear, specific goals and providing feedback (Darling-Hammond & Rustique-Forrester, 2005, p.293). Teachers can use LSA results as a starting point for discussions about instruction suited to struggling students (Shepard, 2010). The allocation of time and resources can be adjusted to better suit students based on test scores (Koretz, 2002). Data, when well-used and

relevant rather than backed by sanctions, can be beneficial to teachers but responses are varied (Louis, Febey & Schroeder, 2005).

Educators also see first-hand what students learn from the assessment process. They relate that excessive testing leads to disengaged students with a poor self-concept (Volante, 2006; Shepard et al., 2011). The hard-nosed, accountability first approach also flattens out linguistic, cultural, and intellectual diversity in classrooms (Luke, 2011). Students quickly learn the rules of the testing system, and eventually strive after only grades, and not understanding (Shepard, 2010). Gains made in performance as a result of early and repeated assessment must be measured against the negative effects on learning and motivation when, “students are taught to measure their success in terms of test scores” (Shepard et al., 2011). However, student reactions are not within the scope of this study as the reactivity of teachers is the main focus, but teachers are certainly affected by their students' feelings and attitudes about LSA. Students are the ones who write LSAs and the results may have momentum that carries beyond these individuals to rebound back on the test-giver.⁷⁶ Teachers have the potential dual misfortune of seeing excessive mandated testing poison their educational well (with students), and then witness themselves being charged with the poisoning.

Figure 4.2: Summary of test design, data and attitudes literature

Topic	Author(s)	Summary statement
	Policy-level considerations	
International testing	Martens, Kerstin, Niemann & Dennis, 2010	Paper asks why thy testing begets testing, and why Germany cares about PISA and the US does not, even in a very accountable age.
	Morgan, 2009	Examines PISA history, its influence on Canada, and the Finnish model of high qualifications and pay-autonomy.

⁷⁶ "To the extent that they [LSAs] are used to measure or reward the performance of educators or schools, they all rely on indirect measurement; that is, the quality of teachers' performance is inferred from students' scores. They rely on high stakes as an incentive for positive changes in practice. They rest on the assumption that the measures employed are sufficient and that estimates of improvement are meaningful. All of these notions are problematic." (Koretz, 2002)

	Volante, 2012	This paper examines provincial policy reactions to international testing. Differing responses seem to depend on relative provincial rankings, tested subjects often had more intense reform initiatives, and wider-ranging geo-political influences.
Accountability function of LSAs	Ben Jaafar & Earl, 2008	A cross-jurisdictional comparison of LSA models in Canada from the perspectives of consequences and the use of data. It includes no small-scale data.
	Espeland & Sauder, 2007	The primary source of the author's reactivity model based on law school reporting of data for 'US News' national rankings. Results show cheating, stretching, and gaming.
	Jacob, 2004	LSA tests in Chicago are studied where increases appear in one school, but not in another. Evidence of gaming or strategic moves by schools.
	Kornhaber, 2004	Paper on the testing regime in the US and how test-based accountability is not an answer to concerns nor can it conquer apparent problems in educational system.
	Linn, 1998	Early but influential paper concludes that there is too much testing and too much emphasis on it. Examines the distinctive 'saw-tooth' pattern of test results which shows reactivity effects.
	Morris, 2011	An encyclopedic run through of LSA testing from design to philosophy, with Canadian examples given as well as guidelines for testing.
	Muriel & Smith, 2011	Examines how qualitative performance measures (LSAs and value-added) can be used to improve teaching and learning. This is an economic view of schooling (outputs, inputs, principal-agent issues, incentives, choice, etc.), yet gaming strategies are mentioned. Concludes that the current system is flawed so better measures are needed.

	Sahlberg, 2010	An examination of accountability practices and their negative effects. He proposes a different responsibility model.
Purposes of testing	Airasian & Madaus, 1983	This study has a strong focus on the instructional link to testing concluding that high stakes use of single measures leads to teaching-to-the-test, and reactivity (citing Campbell).
	Hanushek & Rivkin, 2012	Addresses the use of value-added (VAM) to measure teacher quality and makes note of the effects of measurement error (etc.) but supports the use of VAM.
	Lytton & Pyryt, 1998	This 'effective schools' study in one district (Calgary) tries to explain between-school differences in achievement. Conclude that 50% is attributable to socio-economics, 10% is language (ESL), and 5% is teacher variables.
	Mehrens, 1998	Study on cross-purpose testing, test formats, public reactions to results, stakes, students, etc. Ends with reasonable conclusions and some suggestions on how to improve.
	Popham, 1999	Looks at the issues of cross-purposed and unreliable tests that are used for measuring educational quality. Concludes that the instruments are not meant for (or capable of) this task.
	Volante & Ben Jafaar, 2008	Looking at national (PCAP), provincial and international assessments in terms of student- and school-level improvement. This study looks at theory v. practice in assessment and what tests are intended to do with national and international examples.
	Wenglinsky, 2000	This paper uses 1996 NEAP data from students and teachers to question teacher input/output and their roles in increased achievement.
System-effects on teaching	Haladyna, Bobbit Nolen & Haas, 1991	Looks at test score pollution and how high stakes policies drive poor practices. Includes a chart amended for use in Chapter 3 on ethical/unethical practices.

	Koretz, 2002	Digs into the practice of using tests to evaluate teacher performance. Focus on topics such as high stakes, score inflation, corrupt practices, what are 'meaningful' gains, and negative reactivity in general.
	Luke, 2011	Paper discussing how LSAs flatten diversity in instruction, and degrade cultural variation.
	Mintrop & Sunderman, 2009	Examines how high stakes testing creates the conditions for staff turbulence and compromised teaching standards.
	Nagy, 2000	Canadian study shows teachers do diagnosis with their own instruments. LSA issues are: test design, roll out, and minimal reactivity on external assessment.
Test design, results validity	Cimbricz, 2002	This study is a meta-analysis on teachers' beliefs. It shows there is reactivity and that tests driving instruction. It also shows curriculum narrowing, teaching to the test, and test-like assessment are common.
	Coburn & Talbert, 2006	Paper on cross-jurisdictional differences in ideas about what is good data and how to use it. Includes lots on score inflation, high stakes, PD, collaboration, implementation, etc.
	Darling-Hammond & Rustique-Forrester, 2005	Examines the effects of high stakes in Connecticut, Kentucky, Vermont cases. Authors conclude that policy design and implementation are key, PD is needed and that buy-in can prevent negative reactivity effects.
	Koretz, 2005	This paper examines how both ethical and non-ethical practices can lead to score inflation and undermine the assumptions of the test.
	Rosenkvist, 2010	A wide range of topics are reviewed all on LSA. Themes include results use, high stakes, reporting data, improved instruction, data analysis, and test design.

	Schorr, Firestone & Monfils, 2003	A New Jersey study looking at test design, PD opportunities, and stakes. In many cases, they find, teachers take on strategies in name only, the data have limited reactivity, and PD is test-based and ineffective.
	Skwarchuk, 2004	A good survey of LSAs in the US and then in Canada which also asks Manitoba teachers opinions and reactions to testing. These survey questions were adapted for use in this study.
	Ungerleider, 2003	Examines testing policy from the perspective of teacher engagement, common standards, common misuses of the data, and the need for PD.
	Volante & Cherubini, 2007	This paper examines how to improve assessment practices in classrooms beyond the use of external high school tests. Conclude that PD and university (teachers' college) are possible methods.
	Classroom-level considerations	
Negative consequences	Cizek (2001)	This paper looks at the results of testing which are negative in the majority, but include some unexpected positives.
	Fehr, 2008	Examining university-level financial certification testing, the author finds that teaching to the test an acceptable way to cover lots of content with a self-motivated group. It allows for breadth but little depth.
	Finnigan & Gross, 2007	Chicago study to determine if high stakes increase teacher motivation. The conclusion is positive, motivation and changes in work practices were apparent but, possibly as a result of negative reactivity and support was not forthcoming enough to help teachers with appropriate and effective changes (positive reactivity).
	Koretz & Jennings, 2010	Looks at testing policy issues using real world examples of reactions from teachers coaching and the problem of high stakes.

	Luna & Turner, 2001	Paper looks at the Massachusetts Comprehensive Assessment System tests using interviews. Concludes that high stakes leads to narrowing, test-like preparation, less choice and less creativity.
	Simner, 2000	This study looks at the principles of data reporting, the pressures involved, and specific cases of cheating and reactivity.
	Thomas, 2007	Examines the weaknesses of performance measurement and game-able accountability systems
	Volante, 2007	A compiled list of problems with LSA model and possible solutions with an interesting take on Alberta non-post-secondary.
Positive consequences	Halverson & Thomas, 2007	Paper argues that resources teachers (SSTs) are in-house data experts using 2 case studies related to applying this to instruction at school level.
	Louis, Febey & Schroeder, 2005	Paper looks at how teachers make sense of testing and the effects on practice from a teacher perspective.
	Means, Padilla, Debarger & Bakia, 2009	A US Department of Education commissioned report on how data can and should be used in schools, classrooms, states. Includes 10 case studies in purposively selected districts and has interesting breakdowns of district and school-level supports as well as data interpretation and use criteria for teachers.
	Shepard, 2010	This paper takes a perspective on Halverson's formative feedback model as well as negative effects such as teaching to the test and high stakes. Important factors appear to be staff readiness to analyze data and organizational strength.
	Volante, 2004	The author expresses concerns related to teaching to the test practices, lost instructional time, curriculum narrowing, instructional weakness but does provide some more effective alternatives.

	Volante, 2005	This paper looks at item sampling errors, the improvement of classroom assessment. To be used, LSAs need to be meaningful to teachers.
	Wayman & Stringfield, 2006	Centred on the use of technology to improve use of data, results show some of the benefits of data use and the factors that made it possible.
	Wayman, Spring, Lemke & Lehr, 2012	Includes lots of principal strategies to foster data use, but most are not used by study respondents.
	Young & Kim, 2011	A comprehensive literature review on the uses of data. Includes lots of good detailed information and a stacked bibliography.
Teacher opinions of LSAs	Abrams, 2004	Compares survey results of all high stakes states to Florida results. They are much the same with lots of reactivity effects reported.
	Brown, 2004	Examines New Zealand teachers' ideas about assessment in general following LSA implementation.
	Clarke et al. 2003	Qualitative study in which teachers from Massachusetts, Kansas and Michigan relate their experiences with state LSAs.
	Jones & Egley, 2004	Teachers in Florida relate both positive and negative experiences with current state testing regime.
	Mintrop, 2003	Maryland and Kentucky case studies in sanctions and how educators respond. They lead to turbulence, denial, and little reflection.
	Taylor, Shepard, Kinner and Rosenthal, 2003	An in depth look at the Colorado LSA model and resulting teaching to the test and low morale. The positive side is examined in term of curriculum standards and professional development.
	Volante, Cherubini & Drake, 2008	Examined principals' use of data. Administrators self-rated their skills, and differences were noted between elementary and secondary. Concludes that much PD needed at this level, as well.

4.3 Preliminary hypotheses

H 4-1: Teacher opinions of the data returned to them and of the test domain/structure will influence their willingness to react to the data. Thus, favourable opinions of test design, data clarity, and data return timeliness will have a positive impact on **total reactivity** scores which is synonymous with less neutral reactivity.

H 4-2: Teacher attitudes about the potential utility of test results as measured in five distinct domains (school accountability; student accountability; school improvement; negative test attitudes; and appropriate uses for LSA data) will influence their use of the data. Thus, favourable attitudes about the possible uses of assessment data will have a positive impact **total reactivity** scores which is synonymous with less neutral reactivity.

4.4 Results from surveys – test design and results

The first set of the survey results (4.4, test design and results) will be presented nationally and then province by province since these independent variables (tests and results) differ province to province. The second set will be examined only in the national context (4.7, test attitudes) since attitudes about testing are more philosophical and are not thought to differ as widely based on provincial conditions. The analyses of all these independent variables are left to the end, where trends can be identified and conclusions drawn. These are the variables discussed in this chapter:

Test design and results: These are provincial data and test hypothesis **H 4-1**

- IV1 (independent variable 1) – time of data return
- IV2 – aggregated/disaggregated data
- IV3 – item types
- IV4 – presentation of results
- IV5 – data clarity
- and IV6, ability to act on data

Test attitudes: These are national data and test hypothesis **H 4-2**

- IV7 (independent variable seven) – results use for school accountability
- IV8 – results use for student accountability
- IV9 – results use for school improvement
- IV10 – negative attitudes about testing
- and IV11, appropriate uses for the data

Looking first at the tests and data results, a series of seven questions in the survey asked about the test design and the results that were given back to teachers. None of these items had a common scale for responses – they were all crafted with a range of possible responses gleaned from the literature, discovered through the pilot survey, or known from the researcher's professional experience. The responses were presented as ordinal values – respondents would decide which of several possible choices best described their circumstances. An example is shown to clarify the manner of questions asked and the numerical coding.

A relatively straight-forward question was asked about being able to act on the test results: the question asked was whether it was possible for the respondent and other teachers to use assessment results directly to inform their instruction. Possible answers were: (a) yes, we can act directly; (b) some interpretation is required before acting; (c) we cannot act because we are responsible for analysis; (d) we cannot act because the results are poorly presented; (e) or other (written response). These responses were given values based upon the considered opinion of the researcher about which options would make reactivity most likely. Thus 'act directly' was scored as +1, 'some interpretation' was scored as +0.5, and both 'cannot act' responses was scored as -1. The write in responses were checked and coded as seemed fit. In this case, as in many others, the data were not analyzed solely with regressions based on the final numeric score since deeper analysis is found prior to the regressions (namely in the section that follows).

It is clear that the values assigned were based solely on the judgement of the researcher, but as will be seen in the correlation matrices that follow (and precede each regression) there are frequent **positive correlations** between these variables as coded. While the values were chosen based on informed opinion, the choices made seem to be, by and large, fitting. A copy of the full survey instrument including the scores applied to responses can be found in **Annex 2**.

Note that the following analysis is based on the data shown in the chapter-ending charts and tables section (section **4.11**).

4.4.1 National results

- IV1 – time of data return (see **Figure 4.13**)

Across all ten provinces the proportion of teachers who got timely data (returned in the same school year that tests are administered) was 35% of all respondents. This left 65% getting data the following year, not getting data, or unsure about when they are returned.

- IV2 – aggregated/disaggregated data (see **Figure 4.14**)

There was almost an even split between teachers who got both aggregated and disaggregated data from LSA tests in their respective provinces. A middling 52% of respondents reported getting these detailed sets of data. This means, of course,

that 48% did not get these kinds of data, did not get the data at all, or were unsure about them.

- IV3 – item types (see **Figures 4.15 – 4.19**)

Across the three item types from the survey (selected response, short constructed response and long constructed response) there were generally favourable opinions about the design of LSA tests. 63% of teachers found the items used appropriate. That being understood, there were 34% of teachers who think selected response items are used too frequently and 34% of respondents who believe short constructed response items are not used enough. For longer constructed response there were many respondents who thought they were used too much (12%) but more who thought they were used too little (20%).

- IV4 – presentation of results (see **Figure 4.20**)

This survey question was very basic in that it asked how results were returned or presented to teachers, and all manners of return and presentation were tabulated. By far and away, the national data show that school-based administration presented the results to staff most frequently (59%). In declining proportional order, the next responses were presentation by department heads (14%), a hard copy was provided (6%), results were not seen (5%), the respondent was unsure (5%), local marking was done (3%), the respondents had to ask to see the results (2%), and they were presented by the ministry (2%). Many of these responses were 'write-ins' since it is impossible to consider all the ways that the results might be rolled out. As a result of being write-ins responses, though, it is quite likely that many of the responses are under-reported in these numbers ('provided copy', 'local marking', and 'had to ask' are examples).

- IV5 – data clarity (see **Figure 4.21**)

A large proportion (71%) of respondents indicated that the data were clear and they could understand them. The remaining 29% indicated the results were unclear, partially unclear, or not seen.

- IV6 - ability to act on data (see **Figure 4.22**)

In contrast to the clarity of information, the ability to act on the results data was less overwhelmingly reported. Here, 28% of respondents indicated they could act directly, 44% indicated that some interpretation was required before they could act, and 27% of respondents indicated they could not act on the data. There were also 1% of respondents with other (non-categorized) responses.

- Summary of tests and data result

Many of the response numbers from this section of the survey must look encouraging from the perspective of test-designers, testing policy-makers, and ministry officials in general. Teachers reported overwhelmingly that they support the design of test items, that the data are clear, the presentation of results data was wide-spread and largely uniform, and finally that most teachers get aggregated and disaggregated data to work with. Less encouraging results were that most

teachers do not get timely (same year) results, that there are some test items that are not as widely supported in their current use, and that there are many teachers who either cannot act on the data or who must do interpretation first (which is by no means uniform across the respondents).

How these various factors relate to each other will be examined at the end of this section after provincial data have been presented. Most of the decisions regarding the tests (when they are administered and the items that are included, for example) are beyond the control of the classroom teacher. These policy choices may have an impact on data use, which is reactivity and the dependent variable of this study. The returned data are in some ways crafted by the ministry as well, but more local (school- or division-level) variation exists regarding how the data are presented, what interpretation is done, and whether data sets include aggregated and disaggregated numbers. These data-related factors will also be examined in terms of their effects on data use below.

4.4.2 Provincial results

Alberta

- IV1 – time of data return (see **Figure 4.13**)

More than 70% of Alberta teachers get results the next year, are not sure, or do not get them at all. The national average is lower (65%), but also high. This indicates many teachers do not have at hand the tools they need to make instructional choices based on assessment data in a timely way.

- IV2 – aggregated/disaggregated data (see **Figure 4.14**)

The percentage of Alberta respondents who had both aggregated (grouped) and disaggregated (individualized) data available was 56%. This is better than the national average of 52%.

- IV3 - item types (see **Figures 4.15 – 4.19**)

Over the three item types referenced on the survey, 55% of Alberta teachers thought that certain items were used too much or too little. There was an even split on the use of selected-response and longer constructed-response items (between appropriate and either too much or too little use), but 51% of teachers wanted to see more use of short constructed-response items (40% thought current use appropriate). This is much lower than the national findings that 63% of respondents were satisfied with current usages.

- IV4 – presentation of results (see **Figure 4.20**)

A large majority of Alberta teachers got results from administrators or department heads (more than 83%) with less favourable responses (such as getting a copy of the results without any discussion, not seeing the results, not being sure about results return) much less common (13%).

- IV5 - data clarity (see **Figure 4.21**)

A majority (77%) of respondents from Alberta found the data clear which is better than the national average of 71%.

- IV6 – ability to act on data (see **Figure 4.22**)

The large majority of teachers indicated some interpretation was needed before acting on the results (52%) while lower proportions indicated they could act directly (25%) or could not act on the data (21%).

- Summary of tests and data results

The data available to Alberta teachers appear to be comprehensive, well presented, and thorough. Certainly this should give teachers the information they need to improve instruction guided by these data. Yet not all is positive; there is a reasonable amount of disagreement about the types of items used on the assessments, data are not returned to teachers in a timely way, and some aspects of interpretation seem to fall on teaching staff and their colleagues.

There is no doubt that the proportion of provincial testing in Alberta that is done at the high school level, and specifically as exit exams for grade 12 classes (having high value for final grades and mandatory for all core subject areas) explains part of these results. Grade 12 students are mostly graduated by the time results are returned and the data cannot be used to inform instruction of these particular students. They can and do inform the instruction of some more reactive respondents in a more global sense.

We sit down at the start of the year, and so all the physics teachers will get together, and all the biology guys and all the chemistry guys, and we will take a look at, as a whole, how the students did. They are presented in two different fashions. What the school gets back to them is a document for each individual teacher and for each individual class and then it is a document for the school as a whole.

- AB, High school Science teacher, male

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

- AB, High school Math teacher, female

Alberta respondents have the lowest opinion nationally of the item selection on their LSA tests. They are the only province with more respondents

critical of current usage than supportive of it. It should be said that no test design is likely to please all parties. There are those teachers who much prefer an objective measurement that leads to fewer misinterpretations or re-interpretations of the results. There are also always going to be those educators who believe that restrictive selected response items do not give students the appropriate range of methods to express their knowledge and understanding. No one test can satisfy these competing visions, but certainly it is relevant that more teachers find fault with the current design than those who find it appropriate.

A multiple choice test doesn't allow you any sort of partial marks for partial understanding or partial marks for complete understanding and minor errors. . . It forces, in many cases, a specific problem solving approach when the curriculum is specifically written to encourage a variety of problem solving approaches. Therefore exam does not honour the curriculum despite it being written to test the curriculum.

- **AB, High school Math teacher, female**

The multiple choice, the questions seemed designed to trick the students. I don't think they really did a very accurate job of showing what the students were good at.

- **AB, Elementary English teacher, female**

Leaving some of the interpretation of the data to teachers also poses some issues. It would be nice to think that all teachers are able (let alone willing) to take on these data analysis tasks, but this is not supported by respondents to interview questions. If teachers are not able to do the required analysis, then expectations to use these data are bound to be dashed.

As a physics teacher I am comfortable with statistics and numbers. Our biology teachers, some of them may feel a little bit overwhelmed by the numbers. I know our English teachers are sometimes definitely overwhelmed by the numerical data. But overall, I think the data is laid out fairly well.

- **AB, High school Science teacher, male**

Alberta's rates of reactivity effects (seen in Chapter 3) are generally encouraging with the second highest rating in positive reactivity effects, and a rating in the middle of the pack regarding negative effects.

British Columbia

- IV1 – time of data return

A majority of British Columbia teachers report getting results returned the same year tests are written (54%).

- IV2 – aggregated/disaggregated data

Only 41% of teachers reported having the aggregated and disaggregated results returned to them that allow for individualized instruction and also more global comparisons and planning.

- IV3 - item types

A majority of BC teachers think selected response are over-used on current LSAs (54%) while 42% of respondents think both short constructed response and long constructed response items are used too rarely. Together, there are a slightly higher proportion of respondents who think current uses are appropriate (52%) than those who think current use is not ideal. This proportion of satisfied teachers is lower than the national average of 63%.

- IV4 – presentation of results

Results presentation in British Columbia schools seems to be a concern of teachers. Added together, the responses critical of the current data return methods make up 38% of respondents (this proportion is the largest of all provinces). They reported: (a) only getting a copy with no discussion (7%); (b) not seeing results (7%); (c) not being sure if the results were returned (7%); or (d) having to specifically ask to see the results (a staggering 17% proportion for a written-in response).

- IV5 - data clarity

A majority of teachers found the results to be clear (63%), yet this figure is lower than the national average of 71%.

- IV6 – ability to act on data

Respondents in BC were least likely to report an ability to act on the data in the current LSA regime. Fully 49% said they could not act, 37% said there was interpretation required before acting, and only 12% responded that they could act on provided data. National averages show almost equal 'can act - can't act' numbers (27.5% and 27.3%), while the middle ground of 'some interpretation needed' had 44% of respondents. Teachers in this province feel least able to act on the provided data of to all Canadian jurisdictions.

- Summary of tests and data results

There are some serious concerns reported by teachers about provincial tests used in British Columbia schools regarding the way results are returned to them, and most pressingly related to their ability to act on the data given. In nearly all the categories enumerated in this section (excluding only the provision of same year results data), BC teachers were lower than national averages indicating there are no aspects of testing or data return in which the current system excels.

The ministry has made clear to the public their intention that these assessments be used to enhance instruction and to support teachers in this pursuit. The results above indicate that this goal is far from being met. The most telling deficit appears to be the small proportion of teachers able to take these data and put them to use. If LSA policies are to be well implemented, then this key function has to be addressed somewhere along the line. Numbers this large might be present in a new testing scheme, but the FSAs and provincial exams have been in place for years, and the learning curve should have peaked. More professional development may be a solution, but the existing system may be a terd that defies polishing. The Teachers' Federation is militantly opposed to the policy, the schools and districts cannot seem to improve data reactivity, and the ministry has other concerns - at the start of the 2014 school year teachers were on strike.

Yeah. . . No, we [teachers] are not too happy with our ministry.

- BC, Elementary homeroom teacher, female

Some teachers take it as a point of pride to ignore ministry directives regarding LSAs.

The other point that is striking is the high proportion of teachers who took the time to write in a response about the presentation of results data, namely that they had to seek out these results from administrators or department heads. If 16.7% of respondents chose to write this in, it is almost certainly under-represented in the data set. This betrays a culture not of 'secrecy' regarding data, but of neglecting them. Teachers mostly feel unable to act on the results, and those who do are not provided the results without the initiative to go find them. The LSA system appears from the outside to be dysfunctional and unable to meet its own stated goals.

You know, I don't think we even brought it up [LSA results] for discussion. I might have talked to the other grade 4 teacher and that is it. . . We really haven't used them. . . I guess there are times that if the data really stands out, or if we are not doing that well maybe we would use it.

- BC, Elementary homeroom teacher, female

It is unfortunate that kids have the test in February when I still have March, April, May and June to cover the curriculum. So there are questions on the test that I haven't even covered. . . So that part is frustrating as well, that you can't cover all of the curriculum by February and have them assessed.

- BC, Elementary homeroom teacher, female

Obviously because they are being written for an entire province they are not going to take into account individual students' or even districts' or community's needs or any of the nuance that classroom assessment would. - **BC, Division staff, male**

The lack of policy effectiveness has been noted as British Columbia teachers report the lowest national rate of positive reactivity, and have the highest net tendency toward negative effects (called 'net reactivity' and seen in Chapter 3).

Manitoba

- IV1 – time of data return

Exactly half of respondents from Manitoba report getting same year results, the other half indicate next year results, no results returned, or they are not sure.

- IV2 – aggregated/disaggregated data

The data set returned to teachers from this province does not have the aggregate and disaggregated data to the same level as reported by the national average (30% in Manitoba, 52% nationally). This is the lowest reported provincial rate for this indicator.

- IV3 - item types

A very high proportion of Manitoba teachers are satisfied with the test items used on LSAs. Here, 78% were completely satisfied leaving 22% in either the 'some items used too much' or the 'some items used too little' groups. The national number for satisfied respondents in this area is 63%.

- IV4 – presentation of results

Manitoba respondents show a much wider range of responses for sharing the results data. The number reporting administrative sharing was the lowest in the nation (28%) and negative responses (copy provided but no discussion, do not see, not sure, etc.) are also a much larger proportion than nationally (38% as compared to 18%), and second highest amongst provinces.

- IV5 - data clarity

More than half of teachers report understanding the results data, but 43% do not understand, do not fully understand, or do not see the data. This proportion is higher than national average (30%).

- IV6 – ability to act on data

Manitoba teachers report not having the ability to act on results data at higher rates than national averages (43% to 27%). They also report a slightly higher level of being able to act on the data (33% to 28%). The difference comes in the low proportion of teachers who report the need for teachers to interpret the data (25%) which is nationally the lowest proportion for this metric.

- Summary of tests and data results

There are bright spots and darker areas in looking at the various data collected from Manitoba respondents related to the LSA test design, the way data are returned to them, how clear those data are, and if they can be acted upon directly. In two of these categories, Manitoba is at least average, and in the case of being satisfied with the items types on LSAs, has the highest rating of all 10 provinces. The satisfaction that Manitoba teachers show related to the test design is likely a result of the unique design elements of the early and middle years test used in this province.

It is not necessarily a test, it is called a Provincial Assessment. And what we need to do, we're given, I guess it is kind of like a report card, or a list of 5 different skills or domains that we need to assess the students on by the mid-point of grade 7 and then we send in the data based on what we have collected in class. . .

- MB, Middle years Math teacher, female

Well these [tests] evolved over the years because of feedback from teachers. . . They, basically, they are not sit down tests, paper- pencil tests. What they are now are more an outcome-based assessment where teachers do it and implement it within their daily teaching. So they are able to take kids one on one and do different things with kids. - **MB, Elementary school principal, female**

The school division and the province have told us that they don't collect this data to compare classrooms. It is not like the standardized testing where in the States [the USA] you compare like one school district or one school to another. They are not collecting it for that reason.

- MB, Middle years Math teacher, female

Using tests of this nature would neither remove teacher judgement nor lead to the de-professionalization that is roundly condemned in other jurisdictions. Since these results are locally generated, there is no time lag in producing and seeing the results. Therefore, teachers can interpret them as soon as they collect them.

This burden of analysis can be either an empowering or a daunting proposition, though. Data interpretation in Manitoba seems to fall solely on the shoulders of teachers – there is no report generated by the province that teachers were aware of to guide them in interpreting and using the data. Even data collection seems arbitrary according to some.

And there is a huge divide on that. Like we have five math teachers in our school, and every one has a varying opinion on how that is. So if our school can't really agree, then we go to our school district or division, and we have asked our division to give us more consistency. Because we found some of the schools had all the kids at 'meeting' [expectations] and some schools had all the kids at 'beginning' [to meet expectations] because at their department they had different beliefs of what it should be. . . the province is looking for data that is consistent and this is not consistent.

- MB, Middle years Math teacher, female

What I will say, I think would be another downside, is that they are not always clear what happens with these assessments. We submit them online . . . it is easy, and then they are gone. It is like they're disappeared. Now do we use them? Do we have conversations about them at the school level? Yes, absolutely. . . But I don't get a lot of feedback as an administrator from our division or from the province. So I mean we are doing this. . . but sometimes don't know why we are doing it. **- MB, Elementary school principal, female**

This may be a factor leading the data use rates in Manitoba to be so low. There are many teachers who feel able to act on the data (33%), but many more who note either that they must do the analysis (which is not always a skill possessed by the classroom teacher, or by anyone on staff in small schools), or that they cannot act on the data. The ministry needs to be aware that as a result of these concerns, the LSA program may not perform as it was designed. The results lack clarity for teachers and they do not have the appropriate mix of (especially) aggregated and disaggregated data needed to make classroom level and school levels changes.

High school exit exams in Manitoba make up a smaller number of the tests given, and a correspondingly smaller number of respondents in the survey referenced these grade 12 tests. The nature of these exams is different from the formative, outcome-based reporting of the early and middle years testing, but it can and ideally should inform instruction. Manitoba does less testing than any other province (3 English-related tests and three-math related tests through 12 years of schooling). Whether this amount of testing is appropriate or not is a different topic than that explored in this dissertation since the research question here asks if the data generated are used by teachers. Too often in high school it is said that 'these students are gone now.'

[T]he grade 12 assessment is at the end of the term so the results are rather pointless when it comes to directing the students' future education because in most cases it is complete at this stage (Grade 12 ELA examination is at the end of a term).

– **Anonymous survey comment**

This should not be considered a good reason for not examining and changing instruction based on the data returned. Across all of the relatively small number of assessments done in Manitoba, reactivity effects are tepid as Manitoba sits in the middle of the provinces for positive, negative, and total reactivity (see Chapter 3).

Perhaps the indicator that is the most indicative of a problem of leadership in the LSA program is the manner in which results are presented to staff in Manitoba. The 38% of respondents who reported negative perceptions about the return of data means that the results are not ending up in the hands of teachers, they are not being discussed with the teaching staffs of many schools, and they are not, as the ministry had hoped, 'informing teachers for instructional changes and planning.' Testing policy is seen to have lost its earlier laid out goals, openly discussed implementation, and shared information with all teachers.

And I feel like maybe the first year or two that we actually did it there was a lot more focus on it because it was new and there was a lot of learning and follow up. But I find now. . .we're kind of doing it because we have always done it.

– **MB, Middle years Math teacher, female**

At this point now, after all these years of lots of PD [professional development] and lots of work as school teams, I am not sure that they [the assessments] play that [accountability] role.

– **MB, Elementary school principal, female**

The focus of higher jurisdictions has changed, apparently, and unless some corrective action is taken the policy will have to be re-invented, re-written, or simply written off.

New Brunswick

- IV1 – time of data return (see **Figure 4.13**)

The proportion of New Brunswick teachers getting same year results data is 24%, significantly lower than the national figure of 35%.

- IV2 – aggregated/disaggregated data (see **Figure 4.14**)

New Brunswick respondents report getting the aggregated and disaggregated results they could use to inform instruction at very high levels (65%). This is the second highest proportion reported nationally.

- IV3 - item types (see **Figures 4.15 – 4.19**)

The level of satisfaction shown regarding test item types in New Brunswick is just lower than the national average (59% to 63%). Most satisfaction is reported regarding long selected-response items (77%) while the current use of short constructed response was only considered appropriate by 46% of respondents and selected-response items are considered over-used by 41% of respondents.

- IV4 – presentation of results (see **Figure 4.20**)

New Brunswick teachers reported a very high level in results return and presentation, most of which comes directly from administrators (72%). The negative responses in this province were less than the national average (16% to 18%).

- IV5 - data clarity (see **Figure 4.21**)

A very high proportion of respondents claimed to have a good understanding of the results as presented (81%) which is higher than the national average (71%) and the second highest result nationally.

- IV6 – ability to act on data (see **Figure 4.22**)

A higher proportion of New Brunswick teachers reported the ability to act directly on the data (34%) than national figures (28%), but the numbers for 'can't act' remain relatively high (29%) as do those who report that teachers needed to interpret the data some (37%).

- Summary of tests and data results

New Brunswick respondents had generally favourable things to say about the tests and the data they get from them. In many metrics they are at or near the top of national rankings, and regularly beat national averages. The two areas where the story is not as positive are in terms of the date of returned data, and the item types chosen for the LSAs.

In the areas that seem most closely related to reactive practices, those expected to have a large influence on hypothesis **H4-1** (favourable opinions of test design, data clarity, and data return timeliness will have a positive impact on total reactivity), New Brunswick seems a very good example of doing things right. The data are presented to the staff overwhelmingly by school administrators and discussed as a group. They also include the aggregated and disaggregated information needed to make classroom changes and see the bigger picture for school-wide change and growth. Most importantly, a large majority of teachers agree that they can understand the data, even if somewhat fewer feel prepared to act on it directly.

We have some pretty intricate data for, I believe, the grade 6 and 7 assessments, and then the grade 8 one that they do in grade 9 . . . We get to see where they are provincially, are we on the same standard as them. We also get to see district-wide, are we on the same standard as our district. We get to see, to break it down between teachers that are at our school, and then we get to look at the individuals [students] as well.

- **NB, Middle years homeroom teacher, female**

The teachers that have had the hands-on get it first. We have an opportunity to look at it and really, kind of study it. And then we have a team meeting and literacy and numeracy are the two things at our school we try to make it a team effort to try to boost the scores. – **NB, Middle years homeroom teacher, female**

You get a breakdown, let's pick a topic here, of ahh, right angle triangles or topic that may fit under geometry, and ahh, it'll give me a school-based number of, say they scored 53% at one school and 75 [%] at another school. So to me the information is useless information. - **NB, High school principal, male**

New Brunswick has the highest proportion of positive reactivity reported in Canada (see Chapter 3), but it would be neglectful to say that the system is running as a well-oiled machine. There remains some dissatisfaction about the item types used. This is not necessarily something that can be resolved, but it is yet a concern. Having same year results would also be more informative to allow classroom teachers the opportunity to act on the data with the students they currently see in their classes.

The ELPA (and ELPR) assessments do not appear to be highly regarded as instructional tools. They are similar in many respects to Ontario's OSSLT minimum-competency English exams which may suffer from these same faults. A high school graduation requirement minimum competency exam (MCE) is often only used for its credentialing function. It is not challenging for most students, so the results are ignored – once it is passed, it is forgotten.

Our province is moving towards a more individualized learning plan and unfortunately their assessment has not followed suit. . . Instead of assessing where a students are not. . . To start marking it to say, they are at this grade level, then they have achieved this level. . . – **NB, Middle years homeroom teacher, female**

It is a provincial requirement, so since you are keeping this off the record, or you're keeping this to yourself anyway, I have never seen a student not graduate because they haven't passed the ELPA.

- **NB, High school principal, male**

For the others, results may be examined to see where shortcomings present themselves, but negative reactivity is a common reaction to all 'need to pass' situations. (New Brunswick is the third most negatively reactive province and also has the third highest rate of net neutral effects – see Chapter 3). Setting a low bar, the province has set in place the conditions for schools to do whatever works to get students who struggle over that bar. Perhaps these data are intended to serve the policy-level functions the ministry sets out, namely to 'help policy makers make decisions about curriculum and programming.' According to interview subjects, they certainly do not seem to serve school-level needs.

Newfoundland and Labrador

- IV1 – time of data return

LSA data are returned to teachers in Newfoundland and Labrador the next year in almost all cases (97%), and this is a higher rate than the national figure (65%).

- IV2 – aggregated/disaggregated data

A proportion of 62% of respondents from this province reported getting the aggregated and disaggregated data they can use in schools and classrooms (higher than the national 52% proportion).

- IV3 - item types

Respondents were mostly satisfied with the item types chosen for use on LSAs (79% satisfied to 22% not). The items considered least appropriate were selected response (23% think they are used too much) and short constructed response (20% think they are used too little). All figures show more satisfaction with items than the national data so this province has the highest provincial rating on this category.

- IV4 – presentation of results

The presentation of data was ubiquitous in Newfoundland and Labrador: 67% report getting them from administrators, 33% from department heads and no negative responses were recorded.

- IV5 - data clarity

The reported understanding of results data is just above the national average of 71% at 74%.

- IV6 – ability to act on data

The ability to act on results data is another metric where Newfoundland and Labrador teachers are ahead of the pack. An inability to act on the data is reported by 20% of respondents (nationally this proportion is 27%) and 30% indicate an

ability to act directly (nationally, 28%). The number of respondents who indicate that some interpretation by teachers was needed is 50%, though, and this is higher than the national figure of 44%.

- Summary of tests and data results

Newfoundland and Labrador teachers report the tests and their results data are of a high standard. The data appear thorough and clear, and are well distributed to teaching staff across the school system (100% penetration was reported). They also report a strong sense of satisfaction about the items chosen for the LSAs and had only low levels of dissatisfaction even for the most contentious item type across Canada, the selected-response item. Approval was not, however, uniform:

I've taught all over Canada - I've taught in Alberta, and I've taught in Ontario, and I've taught in BC. I have used their exams and their exam questions and they're just, they're better constructed. I'm not fond of the construction of our test items be they multiple choice or short answer. - **NL, High school Science teacher, male**

What is also true is that teachers in this jurisdiction do not get assessment results back until the following school year. This is not necessarily a fatal flaw, but it is thought that timely results would give teachers a clearer sense of how to act in response. The only area where teachers indicate some trepidation in this section is in the ability to act on the data. They still out-perform national average figures here, but the distinction is surprisingly small considering the strong showing from all other variables considered in this chapter. There is a significant correlation using Spearman's rank order correlation test with a value of 0.208 (which is significant at the $p < .01$ confidence level) between the date of data being returned and the ability to act on those data.

What remains to be seen is if all these encouraging signs related to the testing system in Newfoundland and Labrador lead to what the ministry has asked of the program: 'improved achievement, educational program evaluation, and setting expectations for students.' These goals, as delineated by the ministry, might just as easily be met by negative reactivity practices (more common in high school grades) as by positive ones.

For instance, last year we fell down on the last section and we fell down terribly. . . Now here's the problem, it was the end of the year, it's sunshine, the kids just come off grad, and they have everything else on their mind. So we gave thought to maybe putting [the poorly done unit] earlier in the year. . . Is there some way we could

condense, take some of the extraneous material out, put more emphasis on the mean calculations we fell down on.

- NL, High school Science teacher, male

Teachers in this jurisdiction are the most reactive in Canada (seen in Chapter 3). This does mean, though, that negative reactivity effects (1st in Canada) are even more prominent than the positive effects (3rd in the nation). The relationships will be examined more closely at the end of this chapter.

Nova Scotia

- IV1 – time of data return

More than half (56%) of Nova Scotia teachers get the results returned the same year which is significantly better than the national average (36%).

- IV2 – aggregated/disaggregated data

Respondents report that aggregated and disaggregated data are not provided in most cases (58%). This is 10 % higher than the national average of 48%.

- IV3 - item types

Nova Scotia teachers who give LSAs were slightly more likely to find the test items used appropriate (56%) than not appropriate. Still 51% think selected-response items are over-used, 44% think short constructed-response items are under-utilized, and 30% think the same of long constructed-response items.

- IV4 – presentation of results

A relatively high proportion of respondents from Nova Scotia note that the results are not shared with teachers in satisfactory ways (25%). The national average for this response is 18%. For satisfied teachers, the most common response was that results came from school administration (59%), right on the national average.

- IV5 - data clarity

Respondents from Nova Scotia were less likely to report that the results were easy to understand (65% compared to the national average of 71%).

- IV6 – ability to act on data

The ability to act on the data follows the same general pattern as the national data, but lower proportions for both the 'can act' (26% compared to the national 28%) and the 'can't act' (21% compared to the national 27%) groups. The 'some interpretation' category was above the national average (51% over 44% nationally).

- Summary of tests and data results

There is only one variable in this section that indicates Nova Scotia teachers are provided what they need to react positively to LSA data, and that is same year results provided to 56% of respondents. This is not an overwhelming rate of success in itself, and when you factor in all the more questionable and

negative responses to other variables, Nova Scotia's testing program does not seem to have what it takes to deliver on ministry promises.

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

The Program of Learning Assessment for Nova Scotia (PLANS) branch of the Nova Scotia Department of Education and Early Childhood Development is responsible for provincial testing, and a lot more besides that (this is examined in the introductory chapter). One might well form the impression that the agency has taken on more than they are capable of delivering when examining the results from this section. Nova Scotia respondents were lukewarm about the items chosen for the LSAs, and the relative lack of support does not bode well for teachers using the data. They also report being less than adequately prepared to act on the data.

Teachers report not being provided the aggregated and disaggregated data that would make classroom and school-wide decision making possible. The manner in which results data were shared with teaching staff was criticized by about a quarter of respondents and many of these responses are write-ins (thus are likely found in higher proportions than this). There is also a relatively low proportion of teachers who indicate that the data are easy to understand as they are presented to them. Adding these deficits together, one comes to realize that the teaching staffs in most locations are lacking the direction they need from assessment program directors to act constructively on the data, and yet the tests have proponents, as well.

One of the things in Nova Scotia that they have been careful to do . . . is there are criterion-referenced assessments. And that in itself helps us differ from the standards, from the assessments in the United States where often they don't match standards with the assessments. But for us the assessments and the actual questions are created from the curriculum outcomes and they are created by teachers. **- NS, Division staff, female**

Even so, survey data show that many teachers cannot or do not act on these curriculum-aligned, teacher-created, criterion-referenced assessments.

We noted in Chapter 3 that Nova Scotia teachers are among the least reactive to LSA data overall, but also the only province with 'net' positive results.

Since province-wide assessments of English and math performance are given to Nova Scotia students four times in their twelve years in the public education system, they might expect more bang for these tax-payer bucks.

Ontario

- IV1 – time of data return (see **Figure 4.13**)

Ontario results are very commonly provided the following school year (89% of responses).

- IV2 – aggregated/disaggregated data (see **Figure 4.14**)

At 66% of responses, the data provided to Ontario teachers are provincially the highest rated for providing aggregated and disaggregated data.

- IV3 - item types (see **Figures 4.15 – 4.19**)

Ontario teachers are also among the most satisfied regarding the item types used for LSAs. 75% indicate satisfaction where the national average is 63%. Some teachers would prefer more short constructed response items (28%) and some would prefer fewer selected-response items (17%).

- IV4 – presentation of results (see **Figure 4.20**)

The respondents indicate that results are shared with teachers in a very professional way – only 3.8% responded that there were negative aspects to the presentation of the data from the LSAs. They were presented most commonly by administrators (83%) and department heads (9%).

- IV5 - data clarity (see **Figure 4.21**)

Responses were above the national average in questions related to data clarity. While 79% of teachers report they understand the data, the national figure is 71%.

- IV6 – ability to act on data (see **Figure 4.22**)

A lower proportion of Ontario teachers report an ability to act on the data than is true nationally (19% compared to 28%), and fewer report a lack of ability to do the same (19% compared to 27%). The difference in these values comes from the high proportion of reported 'interpretation before acting' (62%). This is the highest rating for this response nationally (the average is 44%).

- Summary of tests and data results

It is as difficult for policy makers as it is for anyone else to consider the possibility that one can do everything right, and still fail. This is not to say that the LSA program in Ontario is a failure, but there is something of a counter-intuitive finding in this part of the study. By most measures, the data provided for teachers in Ontario are thorough, clear, shared consistently and considered valid by the professionals who are intended to use them. The data are supposed to help 'build assessment capacity in schools' and to guide 'targeted improvement' of instruction. It has been noted that by most measures this seems to be the case.

There is value in knowing where you stand against any benchmark. . . Although not all classroom teachers are aware or can appreciate that these instruments, the rigour of these instruments, are designed and tested. . . by their own colleagues. These are teacher-educators who design these tests. - **ON, Division staff, male**

Why, then, do Ontario teachers report such positive perspectives on the tests and data and then report that they are less confident about acting on these same well-regarded data? There is one aspect of the data that could potentially make acting on them difficult, and that is they are almost always returned in the following school year. If students have moved into another grade, they must then be used to inform the instruction of the teacher one room down. In cases when students have moved to another school (common for grade 6 students - a testing grade in Ontario) the only way these data can inform instruction is with an unusually high level of sharing and trust between the staffs of different schools.

A classroom teacher may or may not see a list of students who were unsuccessful on the test. It depends on the school, and it depends on how much information the principal would like to release . . . Always the percentage of students who were successful or unsuccessful on the test for the school board and province are released to the teachers because those are available on the school website. The classroom teacher? I'm not sure if you were teaching grade 11 you would know who in your class was successful, unsuccessful from grade 10, and certainly the grade 9 Math EQAO. That's rarely shared, individual results, at all.

- **ON, High school English consultant, female**

It bears examination how this uncertainty regarding the use of the data translates into Ontario teachers employing negative reactivity effects very readily (as seen in Chapter 3) and positive reactivity effects less readily. The net reactivity slant in Ontario is the second highest in the nation toward the negative. Only one province reports fewer teachers who 'can act' on the data than Ontario, but British Columbia teachers are generally much less reactive to the data than what is seen in Ontario.

Prince Edward Island

- IV1 – time of data return

A lower than average proportion of PEI teachers report getting results the same year tests are written (29% compared to 36% nationally).

- IV2 – aggregated/disaggregated data

A small majority of respondents reported having access to aggregated and disaggregated data results in PEI (52%). The national average is also 52%.

- IV3 - item types

Most teachers were satisfied with item types used on LSAs in this jurisdiction (73% compared to 63% nationally). Still, 34% would prefer less selected-response, 20 % would like to see more short constructed-response, and 17% would prefer to see more long constructed response-items.

- IV4 – presentation of results

Data are shared mostly commonly by administrators (77%), and negative comments about results sharing were very uncommon (2.9% of responses).

- IV5 - data clarity

The clarity of results data is very high compared to other provinces since 83% of PEI teachers found them easy to understand while the national average is 71%.

- IV6 – ability to act on data

PEI teachers report the highest national proportion of teachers who feel ready to act directly on the data provided (46%). This was the most common response in PEI and much higher than the national average figure of 28%. Still, 29% report the need to do some analysis, and 26% indicate they could not act on the data as presented.

- Summary of tests and data results

Prince Edward Island has a testing program that seems to be considered reliable, consistent and informative by the respondents to the survey. They gave high marks to almost all the aspects of the test and the data returned that were asked. Only with regard to same year results did PEI assessments fail to exceed national averages for those qualities thought to make the data appear more valid and actionable by teachers. It has already been stated that these two factors are significantly correlated.

Still, with all these positive signs, reactivity in Prince Edward Island tends slightly toward the negative (negative reactivity effects are more common than positive effects as seen in Chapter 3). In net reactivity, PEI is the second-highest rated province, behind only Nova Scotia, which alone has a positive net rating.

The situation appears to be similar to that in Ontario where the assessment program appears by all accounts (examined so far in this section) to be striking the rights notes and teachers appear mostly satisfied with the way that testing is done, the end result is not what the ministry had intended.

They are presented loosely . . . Like [administration] kind of holds them - they are sent to office and administration is expected to kind of share them with staff and say this is how we've done. . . Like if I wanted to go to talk to [administration] and really work out the

details I guess I could but you still only have a bare-bones type of result. – **PEI, Elementary homeroom teacher, 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**

When we started it, I was a strong supporter of the assessment. . . I thought, well that's great, now everybody has to teach the same stuff. . . But what I'm finding now is that they start out by saying, now this is just a snapshot, it is not what the child can do all the time, we're not using them for marks. . . I don't find it as effective.
- **PEI, Elementary homeroom teacher, female**

The intention to 'find where students could use assistance and to guide resources accordingly' is one that many provinces aspire to fulfill with their provincial assessment programs. This type of evaluation, standardized and objectively scored, should provide a data set to teachers that confirms what their classroom assessments might have hinted at, or might highlight strengths or weaknesses not yet seen. Either way, if the teachers trust the data and get them presented as is true in PEI, then these results should provide the information necessary to deal appropriately with the needs of all their students. PEI teachers said that they are able to act on the data, and that they think the data are sound, and yet they do not act in the numbers (it is the third-least reactive province) or in the manner (reactivity tends to the negative) that one would expect. Some possible reasons for this were provided by interview respondents:

Most of our curriculum is getting kids to make a connection 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**

There is certainly a discrepancy between say, provincial assessment and how we would assess in the classroom. We would not want, in the classroom, to assess the way they do on the provincial assessment.
- **PEI, K-9 school vice principal, female**

Of course we don't need those assessment results to know which students are, you know, struggling or doing especially well.

- PEI, K-9 school principal, male

Québec

- IV1 – time of data return

A 47% proportion of Québec respondents report same year results being given to them. The others get them the next year, are not sure, or do not see them at all.

- IV2 – aggregated/disaggregated data

Only 38% of respondents got both aggregated and disaggregated data returned to them. This is much lower than the national average of 52%.

- IV3 - item types

Satisfaction with the item types used on the tests was the response chosen by 58% of respondents (compared to a national average of 63%). Teachers less satisfied with the item choices were in favour of more selected-response (unique to this province at 36%), more short constructed response (40%) and fewer long constructed-response (40%).

- IV4 – presentation of results

The return and presentation of results shows administrators and department heads sharing these duties (53% and 13% respectively) with negative responses about the presentation of data relatively low at 17%. The national average for negative responses is 18%. Locally marked tests made up 10% of the responses (second only to Manitoba in this category).

- IV5 - data clarity

This factor was poorly rated in Québec with only 60% of respondents indicating the data were easy to understand. Only Manitoba has a lower rating for this response. The remaining 40% do not understand, do not fully understand, or do not see the data.

- IV6 – ability to act on data

The proportion of teachers who responded they were able to act on the data is higher than the national average (37% compared to 28%). There are lower-than-average numbers for 'some interpretation needed' (also 37%) and 'can't act' (23%). This level for the 'can act' response is behind only PEI nationally.

- Summary of tests and data results

Most indicators from respondents in Québec show that the tests and the data returned are well-suited to teachers making use of them. The only categories where they are poorly rated are in relation to the types of data returned not being specific or general enough, and that the data are not as clear as they might be.

If I have a student who is at a 70 [%] student during the year and he gets a 60 [%] on the exam, I have no way of knowing what did he miss, what was wrong with the exam. I don't get to see the exam. So I can't look and say, you know what, here's what I need to improve. Say all of my students had difficulty in this area of the exam so I need to focus more on that. That type of data I do not get.

- QC, High school English teacher, male

Interviews done with teachers from this province paint a stark picture of the high school assessments. There may be some difference between what assessments provide in low stakes situations (early and middle years) and high stakes one (high schools).

They use words like. . . the student extracts, sort of *general* meaning from the story; the student extracts *insightful* meaning from the story; or the students extracts *highly insightful* meaning from the story. Okay. Like how do you define that? . . . These 16 and 17 year-old kids getting a 6- or 7-page story cold, they've never seen it before. They have to sit down and in an hour read it and analyze it for similes, metaphors, meaning, and they are supposed to get deep meanings out of this in a one hour sitting.

- QC, High school English teacher, male

While negative sentiments about the tests and the results may be in the minority, they do present a different picture than the survey data alone.

There must be some information that teachers feel ready to act upon, as Québec has the most total reactivity of all Canadian provinces (as noted in Chapter 3, this figure is tied with that of Newfoundland and Labrador). The tilt of the reactivity effects is to the negative, though, and Québec has a net reactivity effect that is close to the top nationally of this less encouraging yardstick.

Québec teachers are clearly taking assessments results to heart and acting upon them but the common theme of reactivity effects noted this far, a lack of differentiation between the negative and positive varieties, is a problem. The variables in this section show that teachers agree about many of the choices made in the test design and appreciate the way results are shared. The lack of clarity or incomplete understanding of the results certainly would be an impediment to putting them in action to affect instructional change. Analysis at the end of this chapter will examine this correlation.

Saskatchewan

- IV1 – time of data return

Saskatchewan teachers split 49% of teachers seeing same year results, 51% of teachers seeing results the next year, not sure about the results, or not seeing them.

- IV2 – aggregated/disaggregated data

Respondents reported 57% of the time that results provided aggregated and disaggregated data to them. This is better than the national average of 52%.

- IV3 - item types

Teachers reported test items being appropriate 58% of the time, as compared to over- or under-use of item types 42% of the time. This is fewer satisfied respondents than the national average of 63%. As is common to most jurisdictions, more short constructed-response was preferred (39%) but fewer selected-response items (38%).

- IV4 – presentation of results

Data sharing presentations fell mostly on administrators (62%), as is true everywhere, but relatively high levels of negative response (22%) were recorded in Saskatchewan.

- IV5 - data clarity

Data are clear to fewer respondents than the national sample (62% compared to 71%). This means 38% do not understand, do not fully understand, or did not see the results data.

- IV6 – ability to act on data

This metric looks very much like national averages, with can and can't act responses in close alignment (23% for each in Saskatchewan; 28% and 27% respectively nationally) and 'some interpretation needed' as the most common response (53% in the province, 44% in Canada).

- Summary of tests and data results

Judgements about Saskatchewan's assessment program are guides only to the past since these AfL (Assessment for Learning) program is no longer ministry policy. Lessons learned may be a guide to the future assessments that are surely on the minds of the Minister or Deputy Minister. The references made in the most recent ministry announcements that PISA rankings are falling and need to be addressed certainly point to the idea that more testing will be on the horizon – very few jurisdictions strive for improved test results without doing more tests (save Finland).

This caveat in mind, then, the AfL program was not highly rated by Saskatchewan teachers. Teachers in Saskatchewan were also the least reactive nationally (see Chapter 3). The data were regarded as thorough and the results as timely, but there were concerns about other important aspects of the assessment program. Item types were widely considered unsatisfactory with preferences

much in line with other jurisdictions, asking for more constructed-response and fewer selected-response. The way data were shared was not considered to be helpful or transparent. The most common negative response (with a 12.7% proportion of teachers as a write-in response) was that a copy was given to the teacher, and that was all there was to be done.

I think [our principal] got the results. . . Yes, [I saw the results] but I don't think I did anything as a result of them.

– **SK, Middle years homeroom teacher, female**

In fact, assessment results have not been supported in my school, my division or the ministry. The tests are given, the results are provided in the next school year. AFL assessments have not directed my teaching at all. – **Anonymous survey comment**

Noting in advance that the responses below are from administrators whose perspective is different from teachers in many regards, it can be said that a general neglect of the data was not practiced in all schools.

We would talk a lot about them, we would always come back to them but we tried to be efficient in terms of we didn't want to waste the data. I think it highlighted or it brought things to the front, but we didn't always talk about doing better on the test. It was really, what do our kids need? How can we do better?

– **SK, Elementary school principal, male (a)**

That awareness, I think, created more discussion about language arts, more focus on it. Do I think we did better because of that? There was an accountability piece to it, and yeah, that was probably a positive.

– **SK, Elementary school principal, male (b)**

The two most apparent failings of the AfL program seem to be that the data were not clearly understood by many teachers, and a majority did not feel fully equipped to act on these data. No matter what the other strengths of the program were, it seems unlikely that positive reactive effects would result from teachers who lack understanding of the assessment results and who feel less-than-confident about acting upon them. The fairness of the 'snapshot' test (used to gauge performance only once in the school year) was questioned in interviews:

No, because I don't think it is fair, Every classroom is so different and diverse that I don't think it fair to assume that a small rural

school would have the exact same outcomes or the same abilities of students in it than a classroom in the city that would have 32 students. . . - SK, Middle years homeroom teacher, female

4.5 Correlation analysis – tests and data

The independent variables in this chapter will be examined using Spearman's rank order correlation tests in order to determine:

- What relationships exist between them
- Which relationships complement each other (positive correlations)
- And which relationships are at odds (negative correlations)

Significant relationships will be indicated with an asterisk for significance at the $p < 0.05$ confidence level and two asterisks for significance at the $p < 0.01$ level.

For the six independent variables in this section (see **Table 4.3**), there are many significant correlations apparent. It should not be too surprising that data which are considered to be clear are also quite often returned to teachers and thought to be actionable. None of these correlations (aside from those numbered in negative values) is negative, so the binary values assigned by the researcher appear to align with the perceptions of teacher-respondents. The level of correlation, while significant, is not anywhere near a level where collinearity would be a concern. The most common correlations shown by these data are for **actionable data** (correlated positively four times, three of these at the $p > 0.01$ confidence level) and for the **clarity of data** (three highly significant correlations) variables. The highest levels of correlation are between **types of data** and **data being returned** (0.426) and the **date of data return** and **data clarity** (0.395).

Table 4.3: Spearman's rank order correlation test done with test design and data variables

Correlation matrix - test design and data variables

1. Date of data return	1.000					
2. Types of data returned	0.209**	1.000				
3. Data are returned	0.140*	0.426**	1.000			
4. Clear data	0.395**	0.330**	0.353**	1.000		
5. Actionable data	0.205**	0.142*	0.210**	0.291**	1.000	
6. Test design	-0.055	-0.119*	-0.011	0.027	0.044	1.000

* $p < 0.05$; ** $p < 0.01$

In sum, the relationships in this correlation matrix make clear that the variables in this section are well-aligned, but not so closely as to call into question the model.

4.6 OLS regressions – tests and data results

4.6.1 Regression analysis

Reactivity is the dependent variable in this study, and appears in all analyses below. Yet reactivity (as we have seen in Chapter 3) has both positive and negative effects as defined through the use of the STF Code of Professional Competence (see Annex 1). There are also two other aspects of reactivity derived from these positive and negative results. Total reactivity adds the absolute values of positive and negative effects together to gauge how reactive teachers are to the results data they are presented. Net reactivity subtracts negative effects from positive to determine the overall balance between positive and negative effects. In the regressions that follow, the net reactivity option will not be examined since it is built on mathematically cancelled out values, and thus it would provide an incomplete picture of the data. The other reactivity options will be shown and discussed in this order: positive reactivity; negative reactivity; and total reactivity. Tables show the coefficient in the first row and the *t* statistic in the second row (significant relationships are indicated using an asterisk). Provincial dummies were added to examine variations at this level, and PEI is the control province (it does not have dummy added) for total reactivity, BC for negative reactivity, and MB for positive reactivity.

In examining the survey data it becomes apparent that none of the independent variables examined has a significant relationship to **positive reactivity** effects (see **Table 4.4**). This is somewhat surprising considering the number of interview subjects whose main objections to the testing done in their schools were based on these factors. It can be concluded that since none of these factors has a significant influence on positive reactivity (i.e. the design of the test or the date the data are returned, as examples) they are not effective policy levers to promote the positive use of the data. The realistic interpretation of this fact is to state that an educator who is personally or professionally motivated to use LSA data to improve instructional practices would not be dissuaded by either less-than-ideal test design or less-than-timely results.

None of the provincial dummies shows significance here, so the variation around these findings is certainly not large. Manitoba is the control province in this regression since its positive reactivity scores are closest to the national average for this metric.

Table 4.4: A regression table showing positive reactivity effects in light of independent variables discussed in the preceding sections. Provincial dummies were all added simultaneously to the regression.

Positive reactivity

Date of data return	0.028 (0.31)	0.029 (0.31)
Types of data returned	0.040 (1.31)	0.054 (1.74)
Data are returned	0.156 (1.41)	0.114 (1.00)
Clear data	0.061 (0.59)	0.057 (0.55)
Actionable data	0.083 (0.10)	0.028 (0.33)
Test design	-0.056 (1.96)	-0.038 (1.32)
(provincial dummies) AB		0.292 (1.04)
BC		-0.504 (1.55)
NB		0.313 (1.13)
NL		0.074 (0.23)
NS		-0.050 (0.18)
ON		0.317 (1.06)
PEI		-0.498 (1.73)
QC		0.516 (1.71)
SK		-0.317 (0.96)
Constant	2.497 (24.07)**	2.457 (11.73)**
R²	0.06	0.14
N	298	298

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

An examination of **negative reactivity** highlights some different aspects of these variables (see **Table 4.5**). Before the addition of provincial dummies there are two significant variables at the $p < 0.01$ confidence level and these relate to the returned data's type and timeliness. The relationship for the **date of return** is a positive one and that indicates if teachers report a longer wait for data to get back to them (or that the data are not returned) they are less likely to employ negative reactivity results and, correspondingly, getting timely data actually increases the level of negative reactivity effects, which are instructional tactics specifically used to improve test scores alone. After provincial dummies are added the variable declines in importance to become non-significant.

The second variable that shows significance is the **types of data returned**. Teachers were asked if they were returned classroom, school and/or divisional data from the LSAs their students wrote. Getting all of these varieties of data would include aggregated scores (for schools and divisions) as well as disaggregated scores (for classrooms and students). The relationship between this variable and negative reactivity is a negative one meaning that the more data a teacher is returned, the more likely it is that they will employ negative reactivity techniques. This result means that providing more informative data is conducive to teachers employing 'low road' instructional tactics to improve LSA scores while this same data did not have a significant relationship to positive reactivity effects. For example, an educator is more likely to teach-to-the-test if they have the information showing specifically where scores could or should be improved. An illustration of this effect was provided by an interview subject:

I know the previous year to that we were really bad on quadratic equations . . . We didn't put enough emphasis on it; I know I didn't - I may have done three examples. So this year I nailed the quadratic equation, and we solved that problem.

- NL, **High school Science teacher, male**

Aggregated and disaggregated data sets can (and should be) be interpreted in depth before using them, but a facile analysis may lead teachers to seek gains in LSA scores using negative reactivity approaches ('nailing' specific test content, for example). This finding should provide some support to the policy of several education ministries that providing more information does not always lead to appropriate use of the data since more data can mean more negative reactivity.

Neither of the relationships here are particularly good at explaining the variances in responses regarding negative reactivity prior to adding provincial dummies. Even with two significant variables, the R^2 value remains at 6%.

Table 4.5: A regression table showing negative reactivity effects in light of independent variables discussed in the preceding sections. **It is important to note that since negative reactivity is enumerated in negative integers, a negative coefficient means more negative reactivity effects, not less.**

Negative reactivity

Date of data return	0.257 (2.77)**	0.083 (0.85)
Types of data returned	-0.096 (3.01)**	-0.073 (2.29)*
Data are returned	-0.010 (0.09)	0.119 (1.04)
Clear data	0.035 (0.32)	0.105 (0.99)
Actionable data	0.071 (0.80)	0.073 (0.84)
Test design	0.003 (0.10)	0.009 (0.29)
(provincial dummies) AB		-0.250 (0.75)
MB		0.340 (1.01)
NB		-0.160 (0.48)
NL		-0.440 (1.17)
NS		0.999 (3.02)**
ON		0.148 (0.41)
PEI		-0.436 (1.26)
QC		-0.574 (1.60)
SK		0.371 (0.97)
Constant	-3.093 (28.59)**	-3.227 (11.79)**
R²	0.06	0.18
N	297	297

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

Once provincial dummies are included in the regression (the right-hand column) only the **types of data returned** variable remains significant, and this at a lower confidence level. In examining provincial variations, we should first consider that British Columbia served as the control province in this regression since the average score for negative reactivity in this province was closest to national average. The variation is highly significant for Nova Scotia results, and they show significantly less negative reactivity than the control. As Nova Scotia has the lowest rating for negative reactivity overall, this is not too surprising.

In the final analysis only one independent variable has a significant effect on negative reactivity, and Nova Scotia teachers rated their province to use significantly less negative reactivity strategies related to these variables.

Knowing that Nova Scotia is the only positively reactive province, these variables might go some way to explaining their unique reactivity result. After including the provincial dummies in this regression, a much higher R^2 value of 18% is the result.

Presented next is the **total reactivity** regressions (Table 4.6) which do not stray far from the data seen in the positive and negative effects tables upon which they are based. The **types of data returned** variable remains significant at the $p < 0.01$ confidence level. A positive relationship here shows that as more kinds of data are returned, then it is more likely that a teacher is reactive to it either positively and/or negatively.

Having more data has been shown to inspire some negative reactivity effects, and adding in the potential positive reactivity effects results in a significant finding. It is significant before and after the inclusion of the provincial dummies. Nova Scotia also varies significantly from the control group (PEI in this regression as its average score was closest to the national average) by employing less total reactivity, while Québec indicates more total reactivity based on these factors. Neither of these provincial variations is highly significant, but it bears noticing that including provincial dummies does increase in explanatory power of the regression, demonstrating an increase from accounting for just 6% of the variance in responses to fully 16%.

The most significant aspect of these findings may be that neither the clarity of the returned data nor the perceived ability of teachers to act on them had a significant impact on their reported reactivity. Understanding the data and feeling able to act on the data did not create the conditions necessary for teachers to actually act upon them.

This finding casts some doubt on whether teacher skills/training or organizational strength (which are two possible providers of the resources needed to understand and act on LSA data) have much impact on reactivity.

Table 4.6: Total reactivity effects seen in light of independent variables discussed in the preceding sections.

Total reactivity

Date of data return	-0.242 (1.67)	-0.065 (0.42)
Types of data returned	0.137 (2.72)**	0.132 (2.60)**
Data are returned	0.211 (1.15)	0.044 (0.24)
Clear data	0.016 (0.09)	-0.053 (0.31)
Actionable data	0.007 (0.05)	-0.056 (0.41)
Test design	-0.049 (1.03)	-0.034 (0.70)
(provincial dummies) AB		0.626 (1.46)
BC		-0.377 (0.69)
MB		-0.219 (0.46)
NB		0.609 (1.54)
NL		0.574 (1.25)
NS		-1.023 (2.12)*
ON		0.250 (0.56)
QC		1.206 (2.44)*
SK		-0.591 (1.13)
Constant	5.558 (32.02)**	5.542 (15.13)**
R²	0.06	0.16
N	291	291

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

(Note that teacher training will be seen in Chapter 5 on supports provided for teachers to use LSA data and in Chapter 7 on teacher background factor variables.)

4.6.2 Residual analysis

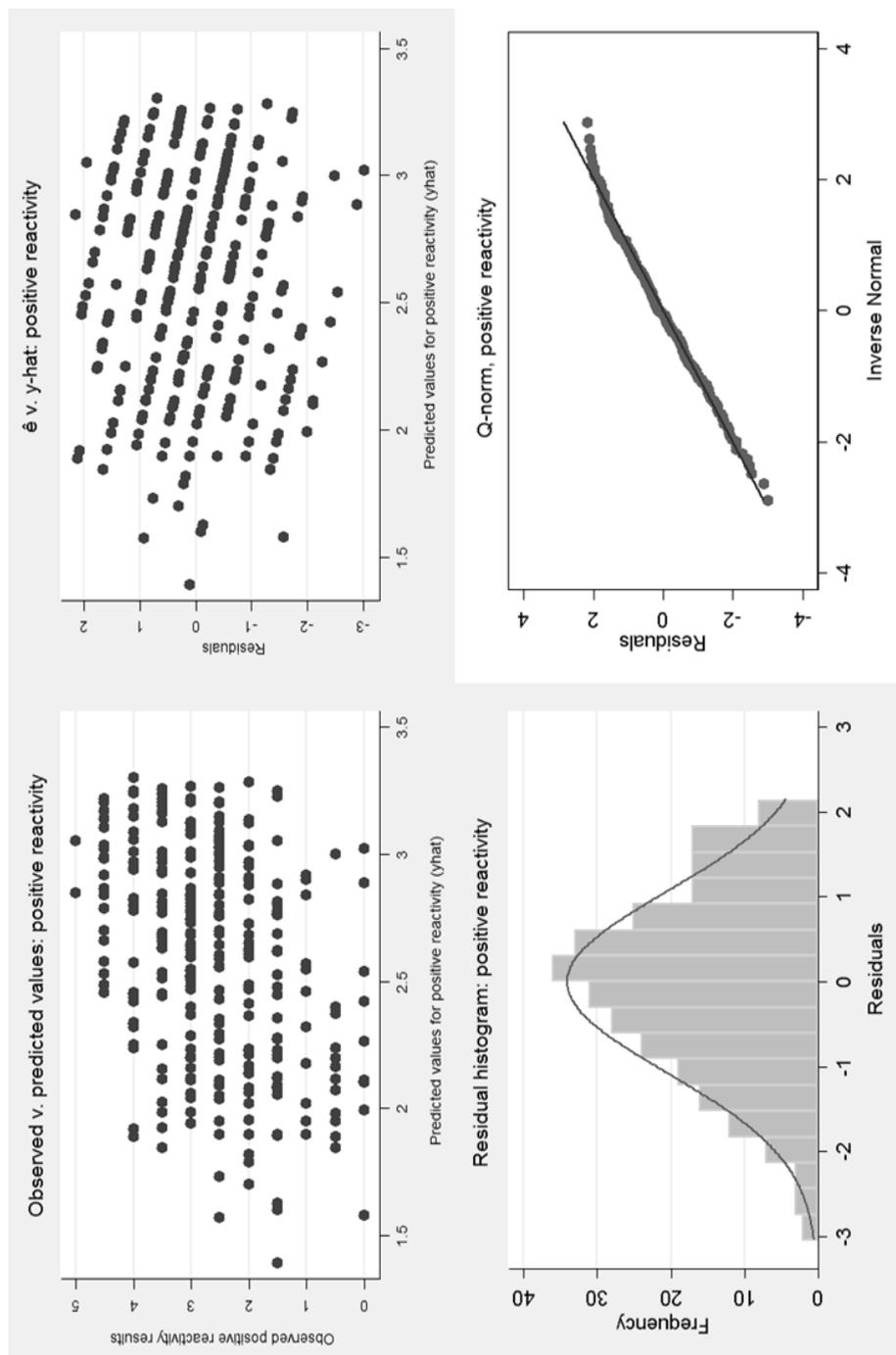
The residuals from these regressions were examined using four different econometric graphing techniques and the results from these analyses were fairly uniform across all three regressions. The results for the **positive reactivity** residual examinations are found in **Figure 4.7**. The 'observed v. predicted values' chart shows a weak positive relation, but as noted from the low R^2 values, there is not a strong relationship between these independent variables and positive reactivity. The bands seen in the graph indicate different reported levels of reactivity (from 0 to 5 by multiples of 0.5). The ' \hat{e} v. \hat{y} ' chart has these same bands, but no clustering or obvious outliers. The presence of clusters or outliers would cast doubt on the use of OLS in the case of this hierarchical data set. The residual histogram has a relatively normal distribution except that the right tail is truncated as a result of discrete variables being used. The closer to the normal distribution the residuals are, the more likely they are to meet the assumption of normally and independently distributed residuals required for hypothesis testing in OLS analysis. Finally, the QQ plot indicates the distribution of the residual is close to normal with some deviation in the right tail. This quantile analysis of the residuals checks the tails of the distributions in particular and some deviation is expected in these data as a result of discrete variables being used. These analyses tend to bear out the rigour of the regression model and help to confirm the findings. (Residual analyses for the other two reactivity types are found in **Figures 4.41 and 4.42**.)

4.7 Results from surveys – test attitudes

Fifteen attitude-related questions were asked in the survey (which were also sub-divided into four categories) to the sample of teachers who themselves give LSAs as well as those who do not. Respondents were asked to 'agree', 'disagree' or 'neither agree nor disagree' with the statements that indicated tests were good for school accountability, good for student accountability, good for school improvement, and if they were useful at all. All these responses relate directly to this chapter's second hypothesis (**H 4-2**) which states: 'favourable attitudes about the possible uses of assessment data will have a positive impact **total reactivity** scores which is synonymous with less neutral reactivity.'

Results for these questions follow, while count data, distribution analyses, and other analyses (where tools are appropriate) are found in the chapter-ending section.

Figure 4.7: Residual analysis for tests and data positive reactivity regressions



The main analyses are based on national averages since the range of responses is quite large and the samples for each province are quite small (in 2 provinces, for test-giving respondents $n = 30$). There is in all cases some discussion of those provinces which deviate most from the national average and in which manner they differ.

The responses in the survey were presented as ordinal values – respondents would decide which of several possible choices best described their circumstances. Examples are shown here to clarify the manner of questions asked and the numerical coding.

The attitude-related questions followed a common scheme: a statement about tests or testing was presented, and respondents could select that they agree, disagree, or neither agree nor disagree. In each case (save the 'negative test attitudes' variable which was scored in the opposite way) answers of 'agree' were scored as +1, and answers of 'disagree' were scored as -1, while neither agree nor disagree was scored as a 0. In this case, as in many others, the data were not analyzed solely on the numeric score since deeper analysis is found prior to the regressions (namely in the section that follows).

So while the values assigned were based solely on the judgement of the researcher, it is important to note that in the correlation matrices that follow (and precede each regression) there are frequent **positive correlations** between these variables as coded. The values were chosen based on informed opinion, and the choices made seem to be, by and large, fitting. A copy of the full survey instrument can be found in the annexes.

Note that the following analysis is based on the data shown in the chapter-ending charts and tables section (section 4.11), reference to which may make clearer the discussion that follows.

4.7.1 IV7 – Using results for school accountability (see Figure 4.23)

Survey question 36 (in **Annex 2**) asks for two responses in relation to this topic. Respondents were asked if they feel that using LSAs as a means of making schools more accountable is appropriate. The rating scale was, as above, three choices: agree, disagree, or neither agree nor disagree. These questions were asked to teachers who give these provincial tests as well as those who do not for comparison purposes.

Both samples agreed on questions regarding the usefulness of LSA tests to gauge school quality and keep education accountable though tracking quantitative quality. More than half of both groups strongly opposed this notion, whereas less than 10% approved of it. Even allowing for one in four teachers who agreed with one of the two prompt statements but not the other, the trend is clearly opposed to 'test-based accountability' (Ehren & Swanborn, 2012; Jones & Egley, 2008;

Hamilton, 2003; Figlio & Getzler, 2002). Interviews went some way to confirm this negative sentiment.

And so we can see overall how our students did on the exam relative to the other schools in the district, other schools in the province, other schools in the area. And so we'll always get a rough 'ranking' idea but, again, without being able to go back and talk about the mechanisms behind it we don't really know what that ranking means. - **AB, High school Science teacher, male**

Now that we have a brand new report card in Manitoba, a provincial report card, and the way we need to report on this report card, I kind of feel now that this provincial assessment mapping that we do is, I won't say is pointless necessarily, but if the whole point is to inform the parents about how their kids are doing, I believe the brand new report card replaces what we used to do.

- **MB, Middle years Math teacher, female**

I don't think that teachers are well trained to administer normative tests and I think that is part of the problem. I think that as a Special Education teacher I have been trained to administer some normative tests so I have a better understanding that adhering to the norms is essential in order to have true results. It is my experience in many different schools. . . No, not at all is it the same between schools. Again, within a town, within a city, and so I can then extend that to a board, there are so many factors that go into an answer. . . It's tricky. I don't think that it's a very reliable test.

- **ON, High school English consultant, female**

And the bottom line here is, like I said, you have this whole 'school success' thing, every year you are supposed to improve. That is not possible to do. It is literally not possible. Eventually, take it to the extreme, eventually you have to have everyone pass with a 100% mark. - **QC, High school English teacher, male**

Accountability, that's quite a, that is quite a, umm, a word, because that an accountability piece, I believe, is the main reason why our government is doing it. Strictly for accountability and strictly as lip service to the general public. . . I just think we, it is something, the accountability, should be left to individual schools and individual teachers. - **PEI, K-9 school principal, male**

These results (that accountability testing is not considered helpful) is not particularly surprising since external oversight is not a popular notion in any profession. Yet support for such policies was forthcoming from some interview respondents. Some interviewed teachers mentioned that teachers' classroom marks had become inflated over the past few years. The competition for university placements and scholarships was often cited as the reason why this was happening. External testing at the high school level provided a check on this trend. Many teachers also cited the importance of collegial professional work with data.

[With no LSAs] All of a sudden the question why is the class average only 70% not 80 [%] has no easy answer. Now the teacher cannot just say the external measures indicate that is the level of the kids. Whereas if that external measure gone, well, maybe next year my kids will all get 80 [%] averages.

- AB, High school Science teacher, male

All the markers are from across the province. It gives a really good chance to align an ensure that we are all looking at the same outcomes. . .to make sure that we are kind of all on the same page.

- NB, Middle years homeroom teacher, female

Because we have the provincial assessments, it allows the opportunity for grade 9 Math teachers to sit in a room two or three times a year and discuss best practices and share best practices and I think it really helps students in the long run.

- PEI, High school Math teacher, male

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

The teaching practices defined as negative reactivity align closely to the conditions of grade 12 exit exams. Variables such as the pressure on teachers, the grade level taught and an awareness of class results are individually correlated against negative reactivity, and the results show significant levels of interaction between these variables. Therefore, more pressure on teachers equates to more

negative reactivity; high grade levels taught results in more negative reactivity; and more awareness of class results also means more negative reactivity. It seems that high stakes testing for students (or for teachers, though this is not the case in Canada) creates the conditions where negative reactivity strategies are more readily practiced, more easily justified, and harder, in many ways, to condemn.

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

This is not to say that there is no value in learning some of those [test-taking] skills . . . But . . . we spent way more time on story-writing than is representative of the amount of time it should get in the curriculum to meet the expectations of the test.

- **AB, Elementary English teacher, female**

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

- **BC, Elementary homeroom teacher, female**

So as a teacher marking it, I thought, wow, I have to make sure that when I instruct . . . this year when I do it again with my students I am going to say, you know, 'Make sure that your, when you divide that rectangle to parts, into four *equal* parts.' I guess stressing various things because of the marking [criteria] that I know would affect their scores on the assessment.

- **PEI, K-9 school vice principal, female**

Whether you agree with or not is another thing, but you've got a lot of kids who are counting on you to get them a scholarship, basically. And you know their situation, they are not going [to university] if they don't get one. - **BC, Division staff, male**

In these cases teachers seem to be painted into a corner by provincial assessment policies and react in the only way that seems fitting. Another sign that the reactivity effects result from the assessment system was the indication that if the testing conditions were removed, the negatively reactive practices would also go.

[My teaching] would change tremendously [without LSAs] because I love the topic. And because I love the topic, what I would do instead, I would do things that are more interesting. I have a lab full of equipment and chemicals and I can only use a certain amount of them. Well, to hell with that; if I had freedom, then I would go down avenues and alleyways that they would enjoy, that I know they would enjoy. . . They might remember that for a lifetime whereas anything you drill into them they forget the next day.

- NL, High school Science teacher, male

In my second year, I went into all the grade 10 homerooms and taught them how to answer a newspaper article or how to write one [which is an item on the OSSLT]. I did not see going into my third year any significant increase even after direct instruction including with the teacher in the room so the teacher could then do this as well with their students afterwards. I didn't see any noticeable increase because, again, that is how to answer a test question. There is no life skill, there is nothing students can take away from that and apply. The question I heard a lot from my staff, again related to the newspaper article, how does this relate to real life? When are they ever going to do this? - **ON, High school English consultant, female**

In the utopia where it was designed, the questions were supposed to reflect the curriculum. But when you have a teacher, it took away the autonomy from the teacher in terms of interpreting the curriculum and implementing it the way it should be. They ended up implementing it the way the test dictated it, the way the questions did. And I felt that was kind of limiting.

- SK, Elementary school principal, male (a)

While provincial samples are relatively small, it is worth mentioning that the provinces least in line with the national data were Québec, Ontario and Newfoundland and Labrador. The trend line across all responses pointed in the same direction, and the lowest possible score (negative 2) was proportionally the most common outcome for respondents from all these provinces, there was a much higher proportion 'on the fence' in Québec and Newfoundland and Labrador (36% and 34% of respondents agreed with one statement but disagreed with the other) while Ontario teachers were the most adamant that LSAs were not appropriate for school accountability (75% of respondents) despite the fact the provincial testing oversight body is named the Educational Quality and Accountability Office.

4.7.2 IV8 - Using results for student accountability (see Figure 4.24)

Three questions were asked about whether teachers felt that using LSAs as a means of holding students accountable is appropriate. The answers were: agree, disagree, or neither agree nor disagree. These questions were asked to both testing and non-testing samples for comparison purposes.

These responses provided a similar result to those about school accountability, perhaps showing that the strongly held negative opinions about test-based accountability are not solely based upon self-interest. The figures for those 'strongly against' have close to a 25% proportion of both groups (which again are in near complete agreement across testing and non-testing samples), while the proportion of respondents who support this function of LSAs for students is less than 7%.

I have some kids that will excel all year, I mean will get in the nineties, get As in math non-stop and then we get to the situational problem and they flunk it. That is why I feel like sometimes those exams do not represent - because of the wording, it is more complicated, it is more adult-like. . . Sometimes I feel like those exams are not fair and they don't reflect the students' abilities because of the way they are worded.

- QC, Middle years homeroom teacher, female

Well I think when you compare schools and you don't take into account the differences and the diversity in each classroom. And you just say, 'this is what it is.' But it is not written down that this [teacher] in this class has three EAL kids or students that are, you know, struggling readers, or that don't get fed breakfast at home or things like that. That is not taken into account. I think that is unfair.

- SK, Middle years homeroom teacher, female

I do think, for all the wrong reasons, that provincial exams at the grade 10, 11, 12 level create a sharper focus for the teachers to pay more attention to those courses and because they know those results are going to be higher stakes and be made public. . . I think people should be focusing on all their courses regardless of whether they're higher stakes or if they're made public, whatever it is, but I think the aspect of accountability, combined with the impact on kids and potentially their futures really does improve the teaching in those courses. - **BC, Division staff, male**

There is a strong positive correlation between the two accountability-based beliefs regarding LSA results (see **Table 4.8**: a correlation value of 0.480, significant at the $p < 0.01$ confidence level). It is not a completely unexpected result that the majority of teachers who do **not** believe that these tests are good at holding schools accountable are the same as those who do **not** believe that they are good at holding students accountable (or those that believe both are true).

Some students, again, students who struggle haven't written a provincial test since [grade] 6 that they may have been exempted from, so they look at the OSSLT because it is a graduation requirement. . . It's tough. They struggle and they don't do as well, and they can't be exempted because it is a provincial requirement.

- ON, High school English consultant, female

If you are going to change it, make it like a standardized test. Make it legit. Make it a standard issue of twenty questions then, maybe don't use it to compare, but if you really want it to be the same for everyone, and then you can choose. . . You can choose to teach to it or you can just choose to have kids do it and see where they all shake down. . . It [the current provincial assessment] is not very specific especially how the data needs to be collected and why.

- MB, Middle years Math teacher, female

And so of the six outcomes that are in the English curriculum one third of the outcomes are ignored by the exam that is worth half of their mark. . . And so you have the diploma exam worth 50% of your mark that only covers one third of the curriculum. And so as an English teacher you want to ensure that their classroom mark accurately reflects at the very least all six of the strands and most certainly includes the two thirds that are not reflected by their diploma exam. So while Math isn't as easily delineated, it is still very similar in that there are still outcomes that you cannot adequately test in a machine-scored way because of the way those outcomes are rendered or experienced.

- AB, High school Math teacher, female

Probably my biggest issue is that they are, umm, used to you know, they are a provincial assessment but they get used right down to the individual students level. . . Really they are just snapshot exams.

- BC, Division staff, male

Well, I think they are just, number one, a very small snapshot of, you know, it is one day. . . We always have to look at the cultural piece around it, you know, we have to look at the test, how kids feel about taking tests, the practice effect. Many of our students aren't used to those formal processes. - **NS, Division staff, female**

Of the provinces that fell least in line with the national average, the balance swings either against testing or in favour of it. For the first trend, British Columbia and Ontario teachers have a proportionally stronger opinion that LSAs should not be used for student accountability. British Columbia teachers have 6% more strongly negative and 8% more mildly negative respondents than the national average. Ontario has 12% more negative and 9% more quite negative respondents than the national levels. The other trend, witnessed in both Manitoba and Newfoundland and Labrador was more support for the student accountability function of testing, but not so much support as to tip the balance in favour of it. Manitoba has 5% fewer strong negative opinions, 4% more mildly positive and 7% more quite positive respondents than the national averages. Newfoundland and Labrador has as many strongly negative opinions, but balancing on the positive side, 2% more mildly positive, 4% more quite positive, and 8% more positive opinions than national levels indicate. BC and Ontario appear to have the most polarized opinions about testing nationally, however Manitoba and Newfoundland and Labrador teachers, while remaining opposed, have (as a group) more nuanced views.

4.7.3 IV9 – Using results for school improvement (see Figures 4.25 – 4.30)

Five questions were asked about whether teachers felt that LSAs results could be used as a tool for school improvement. The choices were: agree, disagree, or neither agree nor disagree. These questions were asked to both testing and non-testing samples for comparison purposes.

Gauging whether LSAs have the potential to improve schools (which would be the likely end result of sustained and wide-spread positively reactive practices), those teachers who give these tests are on balance supportive of the idea. The balance is quite fine as almost as many teachers oppose the idea as support it. Almost one in four teachers has no firm opinion one way or another on this use of LSAs.

It is highly unlikely that they have chosen to hone in this specific chunk of the curriculum. . . So because of that there are questions on every exam where we are just like, 'Oh, well they asked this kind of

question this year. I guess I'll have to spend more time on that the next time I teach the course.' - **AB, High school math teacher, female**

The things I need to assess them [students] on for this, like, forces me to teach maybe the curriculum in a different order or way than I would normally. And I feel like I am kind of jumping around because I need to get them through certain concepts that to me wouldn't necessarily follow [in sequence].

- **MB, Middle years math teacher, female**

I would like to do more project-based learning and I would probably not be focusing as much on the type of assessment that they are doing. - **NB, Middle years homeroom teacher, female**

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

- **MB, Elementary school principal, female**

But, you know, I guess it would be somebody with a bigger perspective than I have who maybe would look at it and say, 'Now wait a second, this is the actual curriculum. These tests are only measuring a small part.' If we pulled back and measured a bunch of different stuff out of the curriculum we would find that those areas have gone downhill because they are not being covered at all.

- **BC, Division staff, male**

There is some uncertainty about the true function of LSAs in schools, and there is some lingering doubt about their ability to improve teaching. These themes emerged clearly from interviews with teachers, administrators and even division-level staff.

Absolutely [there is a lot of variation in information sharing] and two schools in the same town wouldn't necessarily, the staff, have the same information about how the students did.

- **ON, High school English consultant, female**

The school gets the results. You find out like, Sally met expectations reading fiction and she didn't meet expectations reading non-fiction.

But you don't really see what the problems were because you can't see the test. . . It is really hard to interpret it, is what I find.

- PEI, Elementary homeroom teacher, female

It doesn't change the way I teach math. I have always taught math that way. It's just that, let's say, I will choose the things I know that are more tough right at the beginning and spend more time on the tough things than I would. Before I would just go from the beginning and then go all the way through and still, you know, concentrate on certain tough elements. When I get to grade 6 and I know the MELS [provincial] exams are coming, I really, really like concentrate on the tough concepts. So maybe yes, maybe yes it does make the teachers step it up more, maybe.

- QC, Middle years homeroom teacher, female

I wish we had provincial assessments especially in science, language and mathematics. I don't see how we as a province are going to progress when we keep saying literacy and numeracy are the, are the ahhh, quintessential skills sets that we have to have or our students have to have and we don't have any assessment.

- NB, High school principal, male

The relationship between teachers who think LSAs are good for school accountability and those who think they are good for school improvement is apparent and quite strong. The Spearman's correlation in **Table 4.8** shows a value of 0.358 which is significant at the $p < 0.01$ confidence level.

There is a lower proportion of teachers who do not personally give LSAs who believe that these assessments have the potential to improve schools (as compared to teachers who give LSAs). Whether this is from a lack of exposure to the tests or their results data (these teachers are much more likely to never see school-level presentation of the results data) or attitudes from another source is unclear. It is strange is the fact that it will be noted in the data to follow that the non-testing teachers tend to have a more favourable opinion of the tests in general, but as noted above, this opinion does not translate into a belief that they can improve schools.

Regarding this metric, 43.5% of teachers fall into the two 'barely' categories (barely positive, barely negative) or straight 'neutral,' so almost half of the respondents nationally have no strong opinion on this issue. The large numbers seen in the strong positive and strong negative camps (mostly skipping over more moderate response options) indicate that this is also a fairly polarized topic. Interviews bear out the wide range of opinions.

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

The problem with that is, of course, is that it presupposes or has for a basic premise that, is that a significant number of teachers in this province teaching senior subjects don't do valid evaluation in their classrooms and that a three hour exam is a better judge of a student's achievement than my 10 months of evaluation.

- **QC, High school English teacher, male**

We are trying to be data-driven because sometimes that gives you the direction you need.

- **NB, Middle years homeroom teacher, female**

The first year it was rolled out, I wasn't teaching grade 7 Math and then I started teaching it the year after they started using this program. So I wasn't there for all the training, but I really have no idea why it is we are doing all of this. . . And I have the documents and I have scanned through it, but it is still not clear to me what it is we are completing this for.

- **MB, Middle years math teacher, female**

Now it is reported down to the districts and if there are enough students, down to the school level. Those assessments aren't really designed for that. They are really designed to paint in broader strokes. So when you have individual students or individual smaller schools say, 'Wow, we did terrible on the FSA. What's going on?' Well, you know I guess if it is year over year it would be an issue, but, you know, kids not doing well on an exam one particular time, there's all kinds of reasons that happens. - **BC, Division staff, male**

The three provinces with the most variation from national averages on this metric are British Columbia, Newfoundland and Labrador, and Québec. British

Columbia respondents (including both testing and non-testing samples) were strongly against the use of LSAs for school improvement with the highest proportion of teachers falling in the strongly negative category which is true of no other province. The trend line for this metric is therefore negative for BC.

The samples from Québec as well as from Newfoundland and Labrador differ from the national score in that they have more favourable opinions about the potential for large-scale tests to be a tool for school improvement. Québec's distribution of responses has very low numbers in the negative and strongly negative groupings (only 1.6% of respondents in these 2 groups where nationally the figure is 11.8%), but is otherwise fairly similar to national numbers. The trend line is slightly positive for Québec.

This same trend is even more pronounced in Newfoundland and Labrador respondents where there is a higher proportion of strongly positive than strongly negative respondents (true of only one other province – Manitoba). The highest rated responses in this province are: neutral (18%), barely positive (14%), mildly positive (15%), barely negative (12%), and strongly positive (11%). The tilt to the positive attitude is apparent in these numbers.

The national trend line tips just slightly to the positive opinion regarding this use of LSA data. So while there is an over-dispersed but relatively normal distribution of responses on the national sample regarding this use of LSA data, there are provinces at either end of the scale as well, and where support for this practice has more or less traction with educators.

4.7.4 IV10 - Negative attitudes about testing (see Figures 4.31 – 4.36)

Five negative statements were proposed asking for teachers to agree, disagree, or neither. These were a measurement of negative test attitudes and this series of questions was asked to both testing and non-testing samples for comparison purposes.

The results from teachers giving LSAs show that there is a fairly even split between positive and negative opinions of LSAs – the trend line is barely sloped toward positive attitudes. The largest proportion of teachers agreed with some statements and disagreed with others creating a more or less standard distribution.

There are many more teachers on the positive side of this scale from the sample of teachers not giving LSAs and for this group the distribution is certainly skewed toward the positive side. All the negative response categories were outdone by their positive response counterpoints (negative ranging from 10.0% down to 1.8%, while positive ranged from 7.1% up to 17.2%). The trend is also clearly positive.

The best feature, and I think you'd find a lot of science people would agree, is the fact that it validates your assessment strategies throughout the year externally, independently, and I guess objectively. - **AB, High school science teacher, male**

When I taught grade 4 and I had those test results in front of me, then I knew exactly which kids to target with the extra support in certain areas. Sometimes you don't have that information and you gradually figure it out as the year goes along, But when you had that early October mark . . . it was all there, it was very clear.

- **PEI, Elementary homeroom teacher, female**

The best feature of this kind of test is that the kids see how to solve a problem. I find the tests are tough, but it brings them to a higher way of thinking. They have to look at so many elements.

- **QC, Middle years homeroom teacher, female**

I have had a positive experience with them because they are outcome-based and because it aligns with my teaching philosophy. . . I could take pieces of the test, I guess, the assessment and align it more with my philosophy rather than changing my teaching to match a test. . . The kids didn't even really know they were participating in a provincial assessment. . . It was part of our teaching, it was part of everyday and it matched the curriculum.

- **MB, Elementary school principal, female**

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**

Comparing these negative test attitudes to the three independent variables already considered (IV 7 - school accountability, IV 8 - student accountability, and IV 9 - school improvement), the strongest relationship is between *positive test attitudes* and the opinion that LSAs are good for **school improvement** (the correlation value from **Table 4.8** is 0.456, significant at the $p < 0.01$ confidence level). This correlation is nearly twice the value of the correlation to student or school accountability variables. Positive attitudes about the tests clearly are tied to the belief that they can improve schools. Accountability, whether for students or schools, seems to be more of philosophical position, and thus has a weaker connection to positive attitudes about the tests themselves.

In all honesty, I think we're public employees, and I prefer transparency overall. . . I think the majority of the people I have worked with say, you know what, whatever goes on in my classroom ought to withstand public scrutiny.

- AB, High school science teacher, male

The assessments that we do [provincially], of course, are different, and I think that they are productive, like I they give us some good information. But I think that we can get that information from good teaching practices as well, right? So I don't know, I don't know if it is necessary [for accountability] for us to have them.

- MB, Elementary school principal, female

The provinces most out-of-line with national data are British Columbia (where fully 40% of respondents were neutral), Saskatchewan (where polarized opinions are on display with strongly positive and strongly negative opinions both rating higher than the national figures) and Nova Scotia (with a scattered response pattern, the highest rated responses being barely negative, quite positive, barely positive, and quite negative).

4.7.5 IV11 - appropriate uses for the data (see Figures 4.37 – 4.40)

Teachers were also asked about how the data from LSAs might be appropriately used. Those educators who personally give these tests were supportive of a large number of the proposed applications (choosing which parents to contact, or selecting students for specific classes, as examples). Polarization of opinion related to standardised testing is evident in the large bump in the 'all inappropriate' category. Certainly some uses might be considered inappropriate, but any of the options presented could (in the right circumstances) be justified ethically and educationally. Complete dismissal of all proposed uses is thus a relatively extreme position held by 7.4% of the respondents. The majority of respondents indicated that most proposed uses were appropriate (17.2% found all 8 proposed uses appropriate), and the distribution curve for this line of questioning indicates this positive trend.

We can tell from that data [given in October] exactly which classes did what. It does tell me 'did my groups do better on the exam than they did in the year or worse?' and by how much. Typically the results are. . . pretty close to school marks. The variation is anywhere between 2 to 4 percent.

- QC, High school English teacher, male

I'd send that along to the resource teacher in that high school and also to the VP, kind of, who I know looks after that piece to say, 'Well, make sure that teacher knows that Derek needs some support in that area.' But not necessarily with those students who are achieving at grade level. I wouldn't necessarily make contact on them. - **PEI, High school math teacher, male**

It might help if you, year after year, you see that your students don't do well in a certain area . . . but the data should be used to help the students who wrote them. - **AB, Elementary English teacher, female**

Well, assessment drives instruction, right? It should, at least part of it. Umm, and vice versa. So I would think that what they are looking for is . . . growth in the children's achievement. That is the key, it is not whether so much the scores go up or whatever but it is certainly to indicate that the children are able to, umm, you know, that their literacy skills are improving. That's the big part for me.
- **NS, Division staff, female**

From the sample of teachers who give tests, those same respondents who believe that the uses to which LSA data were put were appropriate were also often those who thought they were good at providing accountability and school improvement. There is certainly philosophical correlation between these positions, and this has been borne out in the correlation numbers according to a Spearman's rank correlation test (see **Figure 4.8**).

Teachers who do not give LSAs are just as likely to support the use of results data. As was evident for the sample of test-giving teachers, there are very low numbers shown in the more moderate 'some inappropriate' camp, and a spike of dissent evident in the 'all inappropriate' category. Here, the opposed respondents make up almost 10% of the sample. That said, the trend line remains almost the same as that for the test-giving sample population. The number of respondents in the 'all appropriate' category rises to 18.9%.

I can do that [compare student performance] if I wish. And I can do class to class. I can do school to school if I wish because I also have access now to the performance of, not individual students, but of whole classes of other school in my district, for example, or the entire province. - **AB, High school science teacher, male**

So we usually look at the results when they come in and say, okay so, here's where my kids did poorly, you know, is that the same for

all of us, yes, no, alright, well, what things are you thinking of doing this year in that area . . . so we will talk about things we are going to try. . . - **AB, High school math teacher, female**

It certainly informs our instruction as well, right? . . . I think at the end of it, it's forced teachers to think a little more about asking kids what they think in their subject area instead of, you know, the low-level, recall kind of questions. . . It has forced people to think about how they are evaluating kids. - **ON, High school principal, male**

I wasn't that concerned as a principal about the AfL [Assessment for Learning] scores - I shouldn't say that! I was *very* concerned about the AfL scores but I really felt that. . . If we want to win a football game, we can't just focus on 'winning' because that is not going to help you win. What are the steps that we need to do to make each of our players better on each play? So that maybe we can start to score more points and eventually win a game. I looked at the AfLs the same way. - **SK, Elementary school principal, male (a)**

It is interesting to compare the aggregated results from the **policy-level** test attitudes (independent variables 7 through 10) and the test and data results (IV1 – IV6) examined earlier in this chapter. Teachers were asked about how timely, explicit, clear and trustworthy the data are, and these results were then aggregated and correlated with those policy purposes they believe the data might effectively be used (for school or student accountability, and school improvement, again with these results aggregated). The small adjusted R^2 value (0.0351) but significant t score (4.18) shows that the aggregated values of these two lines of inquiry are not very strongly correlated.

There is a stronger correlation between **student-level** purposes (as opposed to policy-level and seen above as IV11 – appropriate uses for the data) and the first lines of inquiry in the chapter, tests and data (IV1 – IV6). The t score is 7.04 and the adjusted R^2 is 0.1014 in this correlation. This difference may be ascribed to the fact that policy-level opinions are just that: opinions, whereas student-level purposes are related directly to a teacher's day-to-day professional life. Believing that LSAs don't necessarily provide good school accountability does not mean a respondent think that the data cannot be used to contact parents, talk with an instructional coach, or select students for specific classes or programs. These are practical choices that are made at the school, and they are often based, at least in part, on LSA data. A teacher who is more likely to see value in the tests, to trust the validity of the results, and who also finds that they are timely and

accurate is a teacher more likely to think these data are helpful in making informed decisions about students and teaching.

4.8 Correlation analysis – test attitudes

As in the previous correlation matrix, the independent variables in this section will be examined using Spearman's rank order correlation tests to determine relationships that exist between them, positive or negative. Significant relationships will be indicated with an asterisk for significance at the $p < 0.05$ confidence level and two asterisks for significance at the $p < 0.01$ confidence level. It is worth noting for the table below that negative test attitudes responses (this being a 'negative' outcome in a group of positive attitudes) were assigned negative numerical values, and thus the correlations show proper alignment with the other variables.

The results thus far related to the independent variables have not been unexpected. There are also highly significant correlations evident between, in particular, the two accountability and the school improvement variables. These correlations are both common and strongly significant (see **Table 4.8**). These same variables are less closely tied to appropriate uses for test data, and negatively correlated with negative attitudes about testing. This result, which affirms that all relationships are complementary, confirms to some degree the binary values assigned by the researcher to survey responses. The highest levels of correlation are between school and student accountability (0.480) and the negative relationship between school improvement and negative test attitudes (-0.456). It is natural to think that those teachers who feel the test have more appropriate uses are less likely to be critical of the assessments.

Table 4.8: Spearman's rank order correlation test done with test attitude variables

Correlation matrix - test 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

* $p < 0.05$; ** $p < 0.01$

4.9 OLS regression analysis – test attitudes

4.9.1 Regression analysis

Provincial dummies were added to examine variations at this level. Manitoba is the control province (it does not have dummy added) for positive reactivity, British Columbia for negative reactivity, and Prince Edward Island for total reactivity.

Examining the data regarding test attitudes, there are some significant and relatively strong relationships with **positive reactivity** (see **Table 4.9**). Before provincial dummies are included, four of the five attitudes variables have a significant correlation, and three of these four are at the $p < 0.01$ confidence level. These attitude variables account for 18% of the variance in responses regarding positive reactivity. After the provincial dummies are added, three of these variables remain significant, two of them at the $p < 0.01$ confidence level, and the R^2 value increases to 23%.

Starting with the **school improvement** variable, it is clear that for those teachers who feel that LSAs are an effective tool for improving schools, there is a strong and significant increase in the use of positive reactivity strategies. It stands to reason that those teachers who believe that instruction at schools can be improved by examining the results of LSA tests are also those teachers who do the more involved work implied in positively reactive instructional change. This relationship is strongly significant before and after the addition of provincial dummies.

Closely related to school improvement and with similar strong ties to positive reactivity before and after dummies are added is the **appropriate uses** variable. Those respondent teachers who thought that LSA data have multiple potential uses for teachers and schools were more inclined towards positive reactivity. As mentioned above, employing data-informed methods may also mean more work for teachers, and those teachers who think that data could be used in many ways (which entails more work for these same teachers) are the ones who opt for positive reactivity. By contrast, teachers who reported supporting **student accountability** (which is a weaker and negative correlation only significant before provincial dummies are added) may therefore view work being downloaded onto students and away from the education professional as a positive, 'real-world' instructional strategy.

The relationships between both school improvement and appropriate uses are stronger than the correlation for student accountability. Along the same line, the **school accountability** variable does not even rate as significant even though it is very strongly correlated with student accountability (as seen above in **Table 4.8**).

Table 4.9: Regression tables showing test attitude variables compared to positive reactivity effects.

Positive reactivity

School accountability	0.064 (1.24)	0.101 (1.93)
Student accountability	-0.072 (2.00)*	-0.066 (1.78)
School improvement	0.090 (3.54)**	0.077 (2.95)**
Negative test attitudes	-0.083 (3.29)**	-0.066 (2.58)*
Appropriate uses for data	0.044 (3.55)**	0.039 (3.18)**
(provincial dummies) AB		0.658 (2.54)*
BC		-0.242 (0.87)
NB		0.484 (1.91)
NL		0.337 (1.17)
NS		0.274 (1.02)
ON		0.442 (1.67)
PEI		0.041 (0.17)
QC		0.312 (1.09)
SK		-0.057 (0.22)
Constant	2.484 (28.42)**	2.312 (11.52)**
R²	0.18	0.23
N	344	344

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

A somewhat weaker correlation exists between **negative test attitudes** and positive reactivity. Not surprisingly, this relationship is negative, so that those respondents with the fewest reported negative attitudes about LSA tests and policies were more inclined toward positive reactivity. This relationship is also one that makes intuitive sense.

The interesting and significant results here indicate that teachers who report that tests are either unreliable sources of data or that results data are simply filed away are certainly less likely to use these data to improve their instruction. Those respondent teachers who indicated that LSAs were an effective tool for promoting student and parent educational accountability were also somewhat less likely to employ positive reactivity strategies. It might be interpreted as meaning that teachers who want to hold students accountable for their learning are somewhat less likely to hold themselves accountable for instructional improvement based on LSA results. This is seen in the literature as beliefs about an external 'locus of control' (Lytton & Pyryt, 1998 and Schildkamp & Kuiper, 2010). This term describes when teachers feel unable to influence student achievement as a result of factors outside of their control (such as parenting).

Looking at the provincial dummies, there does not appear to be much provincial variation from the control group. Only Alberta has a weak positive correlation indicating somewhat more positive reactivity than the control group, while all other provinces have no significant variances. The independent variables remain highly significant, two of them at the $p < 0.01$ confidence level, and the R^2 increases to 23%.

Negative reactivity correlations (see **Table 4.10**) indicate no significant relationships for the independent variables once provincial dummies are added, and only one weak negative relationship between the **student accountability** variable and negative reactivity prior to the inclusion of the additional variables. This result gives limited support to the described positive reactivity relationship in that respondents favouring student accountability options are less inclined towards positive reactivity but do tend to use more negative reactivity strategies. More striking here is the number of highly significant provincial variations from the control group. Saskatchewan, Nova Scotia and Manitoba all have highly significant positive correlations with the national data (British Columbia is the control for this data).

The inclusion of the provincial dummies increases the R^2 value of this regression from 4% to 20%, yet it is difficult to draw policy-based conclusions from the telling level of variance in scores being explained by these variables except in light of **provincial variations**, which are strongly significant. We can say with certainty that Nova Scotia, Manitoba and Saskatchewan have results that indicate significantly less use of negative reactivity strategies in light of attitudes variables.

Table 4.10: Regression tables showing test attitude variables compared to negative reactivity effects. **It is important to note that since negative reactivity is enumerated in negative integers, a negative coefficient means more negative reactivity effects, not less.**

Negative reactivity

School accountability	0.110 (1.81)	0.033 (0.58)
Student accountability	-0.083 (2.00)*	-0.074 (1.87)
School improvement	-0.024 (0.83)	-0.033 (1.15)
Negative test attitudes	0.053 (1.85)	0.016 (0.60)
Appropriate uses for data	0.002 (0.12)	-0.010 (0.75)
(provincial dummies) AB		-0.107 (0.39)
MB		0.790 (2.67)**
NB		0.191 (0.69)
NL		-0.147 (0.47)
NS		1.083 (3.78)**
ON		0.554 (1.91)
PEI		-0.375 (1.41)
QC		0.048 (0.16)
SK		1.049 (3.78)**
Constant	-3.069 (30.13)**	-3.415 (15.78)**
R²	0.04	0.20
N	347	347

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

Being a metric of positive and negative effects combined, the **total reactivity regression** (Table 4.11) indicates several significant relationships. They are not nearly as strong as those noted for positive reactivity. The provincial data, on the other hand, maintain the same levels of significance for the same provinces noted in the negative reactivity regression.

Based on strong positive reactivity effects, the **school improvement** and **appropriate uses** variables both are significant at a lower level of significance for total reactivity ($p < 0.05$). Strangely, the strongest correlation prior to the addition of provincial dummies is the negative relationship between **negative test attitudes** and **total reactivity**, but it is not significant when the dummies are added. This negative correlation indicates that those teachers with fewer negative attitudes about tests are more reactive (in both positive and negative senses of the word) to the results data. So those respondents who felt the LSAs were an imposition on their classroom teaching and were least trusting of the validity of the results might also be seen to be unlikely to use the results.

The provincial variations here seem to align well with both the previous regressions and the data from Chapter 3 on total reactivity effects. The three provinces with significant negative variation from the control group are three of the four *least* reactive provinces, and thus the negative relationships with total reactivity in this table bear out that this may be largely rooted in attitude-based factors.

An examination of the attitudes variables from this study has highlighted some interesting and significant findings. **Accountability** is quite commonly the stated rationale for LSA testing used by teachers, and yet teachers who responded most favourably to accountability questions on the survey were also very unlikely to use the data to inform their practice. School accountability has no significant correlations with data use, and student accountability has only a fleeting and weak correlation with the use of more negative reactivity effects. Teachers who indicated having strong negative feelings about the LSA tests or policies were also less inclined to use the data.

Strong correlations to positive effects, and by extension, total effects, come from two attitudes variables (**appropriate uses** and **school improvement**). These two variables are correlated strongly with positive reactivity but not significantly correlated with negative reactivity effects. The R^2 values for these regressions are also quite large: 23% for positive reactivity, 22% for total reactivity, and 20% for negative reactivity. What is true of the negative reactivity result in particular is that the high value is almost exclusively the result of provincial variations (these being the only significant correlations) while attitudes factors are significant for positive and total effects.

The provincial dummies also seem to bear out what has already been noted about different reactions from the provinces to LSA data.

Table 4.11: Regression tables showing test attitudes variables compared to total reactivity effects.

Total reactivity

School accountability	-0.047 (0.51)	0.064 (0.72)
Student accountability	0.005 (0.08)	0.005 (0.08)
School improvement	0.115 (2.60)**	0.113 (2.53)*
Negative test attitudes	-0.138 (3.16)**	-0.084 (1.94)
Appropriate uses for data	0.048 (2.21)*	0.053 (2.55)*
(provincial dummies) AB		0.324 (0.84)
BC		-0.618 (1.48)
MB		-1.165 (2.77)**
NB		-0.131 (0.36)
NL		0.067 (0.15)
NS		-1.271 (3.16)**
ON		-0.536 (1.37)
QC		-0.155 (0.35)
SK		-1.572 (4.09)**
Constant	5.539 (36.25)**	6.130 (21.69)**
R²	0.12	0.22
N	339	339

Tables show regression coefficients; t-values of the regression coefficients are bracketed; * p<0.05; ** p<0.01

Nova Scotia is significantly less negatively reactive than the controls, and three provinces are significantly less inclined to use the data at all (Nova Scotia again, but also Manitoba and Saskatchewan).

4.9.2 Residual analysis

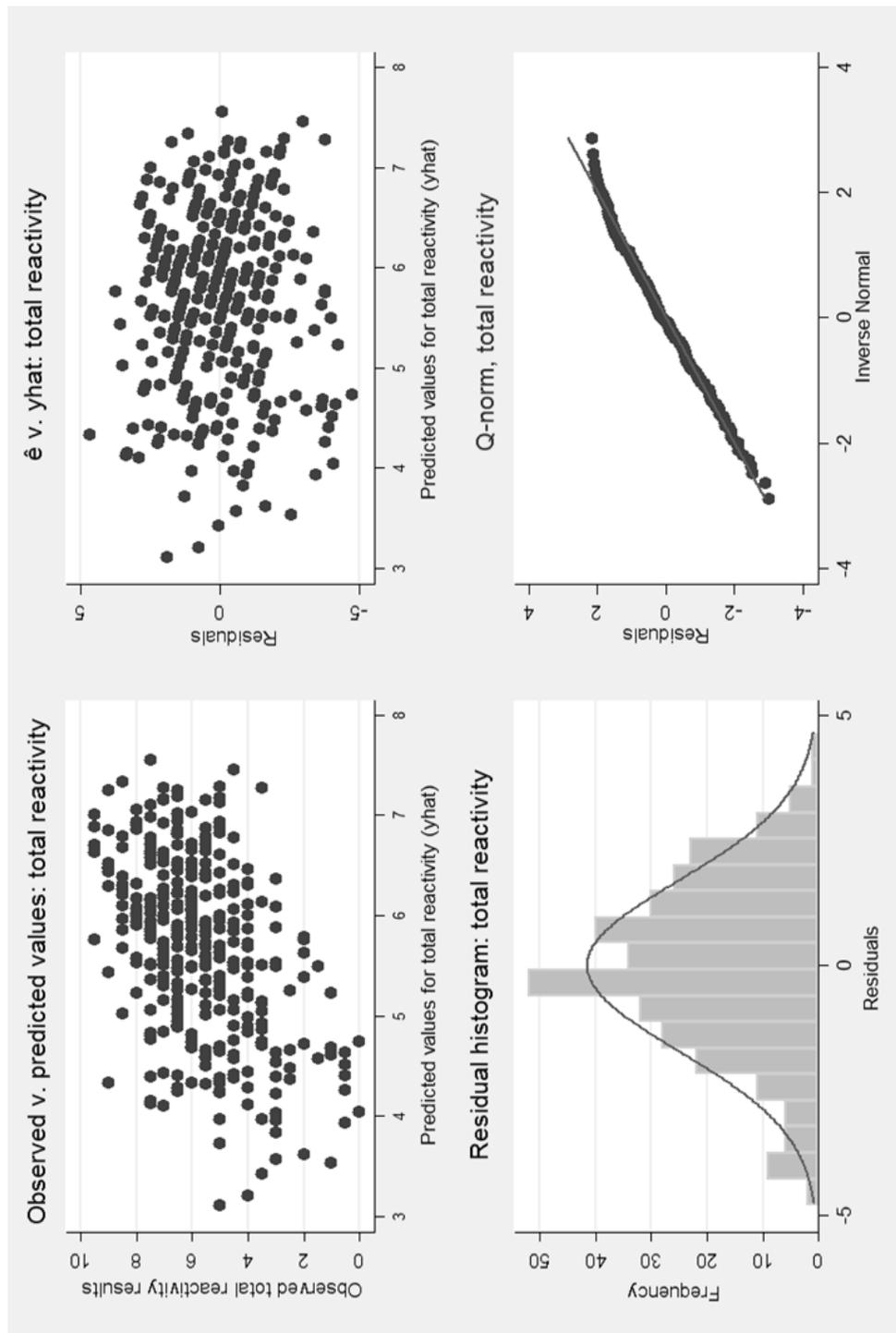
The residuals from these regressions were examined with four different econometric graphing techniques and the results from these analyses were fairly uniform across all three regressions. The results for the **total reactivity** residual examinations are found in **Figure 4.12**. Total reactivity is a comprehensive metric since it includes both the positive and negative reactivity results. The 'observed v. predicted values' chart shows a positive linear trend. Certainly this is not a strong visible trend, but it is better than seeing no linear trend that may indicate a low correlation. The bands seen in the graph indicate different reported levels of reactivity (from 0 to 10 by multiples of 0.5).

The ' $\hat{\epsilon}$ v. \hat{y} ' chart has these same bands, but also shows no clustering or obvious outliers. The presence of these kinds of aberrations would indicate the use of OLS in the case of this hierarchical data set might not be appropriate.

The residual histogram has a relatively normal distribution excluding a few higher or lower bars. This distribution is close to normality, and thus the residuals are more likely to meet the assumption of normally and independently distributed residuals required for hypothesis testing in OLS analysis.

Finally, the QQ plot indicates a small deviation from the normal distribution in the right tail where fewer responses were recorded in the sample. In all, these analyses tend to support the conclusions from this chapter regarding the relative strength of the correlations and bear out the rigour of the regression model to validate some of what has been concluded in the text. (Residual analyses for the other two reactivity types and variables from this section are found in **Figures 4.43 and 4.44**)

Figure 4.12: Residual analysis for test attitudes total reactivity regressions



4.10 Conclusions

Neither of the hypotheses from this chapter⁷⁷ when tested with these survey data have proven to play out exactly as expected, but both have been found to have some merit. The hypothesis that opinions about test design, test data, and getting suitable data would impact reactivity effects has been borne out regarding only the provision of disaggregated (student-specific scores) as well as aggregated (class, school, and divisional data).

Opinions about the design of tests such as item choice are not as revealing. Nor are opinions about the clarity of the data, the perceived ability to act on the data, the timeliness of data return, or the presentation of the data as important in terms of reactivity. This is a surprise since some of the main impediments to their use cited by survey comments and interview subjects were these same factors. In the end, it is the quality and thoroughness of the provided data that made a difference to teachers using LSA results to improve their classroom instructional practices. There being only one significant correlation, these variables did not go very far in explaining the variance in survey responses prior to the addition of provincial dummies – the R^2 value is just 6% for all varieties of reactivity effects examined (positive, negative, and total). After including the provincial dummies these values increased respectively to 14%, 18% and 16%, which indicates that there is strong amount of variance in reactive effects related to test and data.

The second hypothesis, which stated opinions about testing in general and ideas about how they can be used, were shown to have a much stronger connection to reactivity effects with the exception of the accountability functions (both for schools and for students). Having some general consensus between the ministries and teachers seems to be needed to change practices regarding large-scale provincial tests and how the data might be used to improve teaching. This is not currently happening in all cases, but where opinions about testing are positive, the reactivity type most affected is positive reactivity. An unexpected finding was that positive opinions of tests (or at least fewer negative opinions) and supporting several uses for the data is a driver for positive reactivity while negative opinions

⁷⁷ **H 4-1:** Teacher opinions of the data returned to them and of the test domain/structure will influence their willingness to react to the data. Thus, favourable opinions of test design, data clarity, and data return timeliness will have a positive impact on **total reactivity** scores which is synonymous with less neutral reactivity. **H 4-2:** Teacher attitudes about the potential utility of test results as measured in five distinct domains (school accountability; student accountability; school improvement; negative attitudes; and appropriate uses for LSA data) will influence their use of the data. Thus favourable attitudes about the possible uses of assessment data will have a positive impact **total reactivity** scores which is synonymous with less neutral reactivity.

regarding these same aspects do not affect negative reactivity significantly. The link between accountability, the primary rationale for LSA testing according to most ministries and most supportive teachers, does not have any significant impacts on the use of data. Only the number of uses for data considered appropriate and the idea that testing can lead to school improvement are tied to reactivity effects (mostly positive). These factors are instrumental, alongside provincial differences, in explaining the variance in survey responses. The R^2 values for these multivariate regressions are 23% for positive reactivity, 22% for total reactivity, and 20% when looking at negative reactivity effects.

Successful implementation of large-scale assessment policy is dependent on an active connection between staff at the education ministries who both write the assessment policy and set the expectations for using the data (at several levels) all the way down to schools where teachers administer the tests and get back data that may be of use to them in changing their instructional practices. The links in the chain that can help it hold firm despite individual resistance would be both school-based administrators and division-level staff who pass along information, carry out appropriate follow up checks, and interpret ministry goals based on their own sets of beliefs. Not until the policy expectations are clear to everyone and consistent in their implementation can ambitious goals be met. The results in this section are one perspective on of how much fidelity is retained from the ministry's directives to where they are ultimately put into practice when a teacher utilizes (or does not utilize) data-informed decision-making (as examined in Means, Padilla, Debarger & Bakia, 2009; Schildkamp & Kuiper, 2010; and Wayman, Cho, Jimerson & Spikes, 2012).

Having a thorough and complete set of data helps somewhat in this pursuit, but it is clearly more telling to have teachers who hold positive opinions about how data can and should be used. Years ago, the RAND Change Agent study (McLaughlin, 1990) stated that it is a bad policy that works under the premise that 'belief follows practice' – that if a given policy is implemented and its worth becomes clear to its users, then somehow initial resistance reactions and negative attitudes will change. This chapter seems to validate that claim in that only if there is belief in the policy from the outset, that is, if it is supported by teachers, will their practices change. Since these attitude factors account for fully 23% of the variance in positive reactivity responses, the importance of this should not be understated.

4.11 Charts and tables

Figure 4.13: This chart shows the percentage of respondent teachers in each province who were getting timely (same year) results as compared to those getting results the next year, were unsure of when, or did not see the results at all.

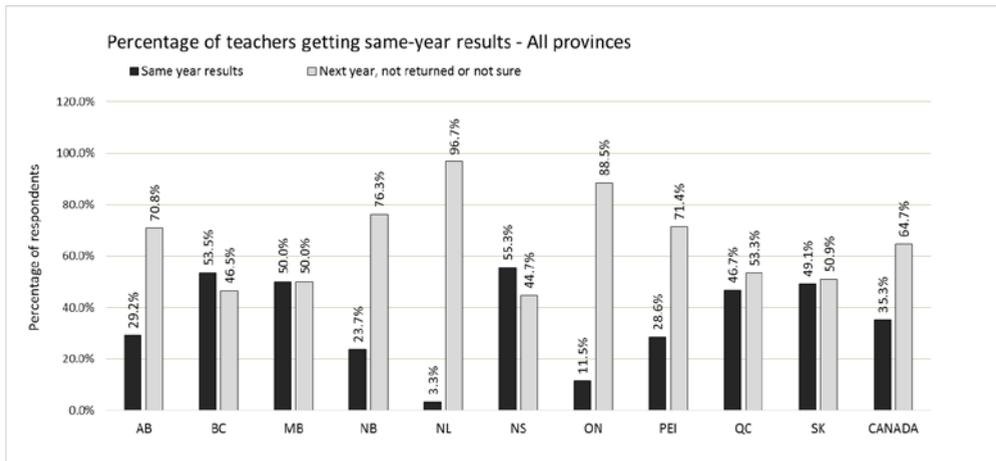


Figure 4.14: This graph shows the numbers of teachers who responded they receive data aggregated (by school division or by school) and disaggregated (by classroom and by student). Responses are grouped to contrast the sole positive response (data are given) and the three negative responses (data are not given, data are not seen, or the respondent is unsure).

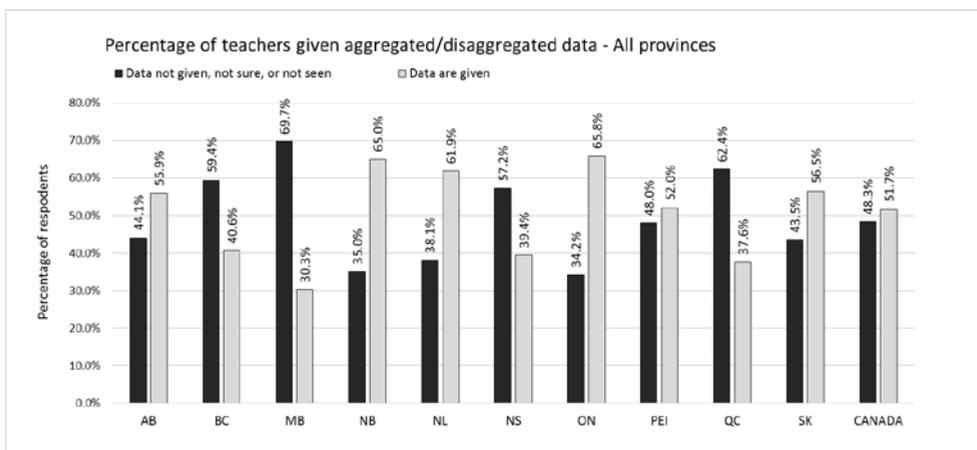


Figure 4.15: Overall results for teacher ratings (item types were designated as being used too much, used too little, or used an appropriate amount) for item types used in their provincial LSAs. These data are broken down by item type below.

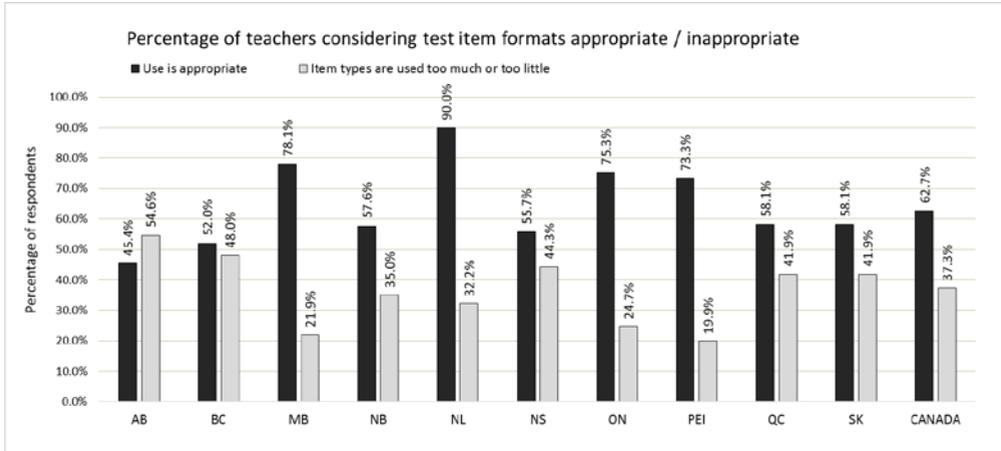


Figure 4.16: Item-usage rating for selected-response items like multiple choice, true/false or matching questions.

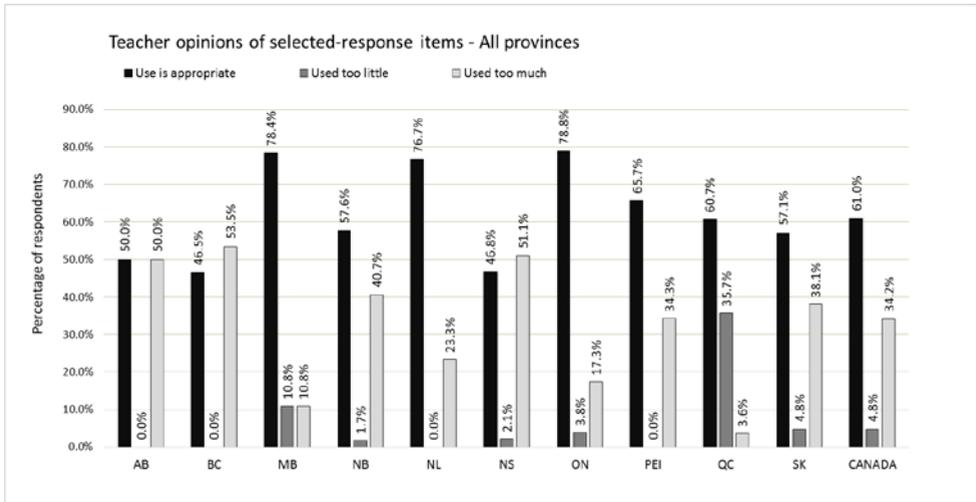


Figure 4.17: Item-usage ratings for short constructed-response items like fill-in-the-blanks, definitions, numerical response, or short answer questions.

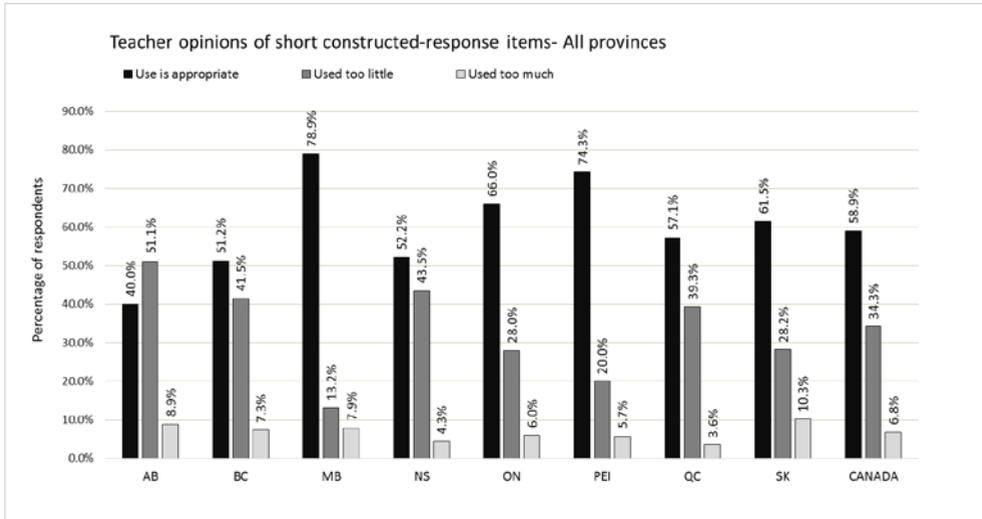


Figure 4.18: Item-usage ratings for long constructed-response items like paragraphs, word problems, or essays.

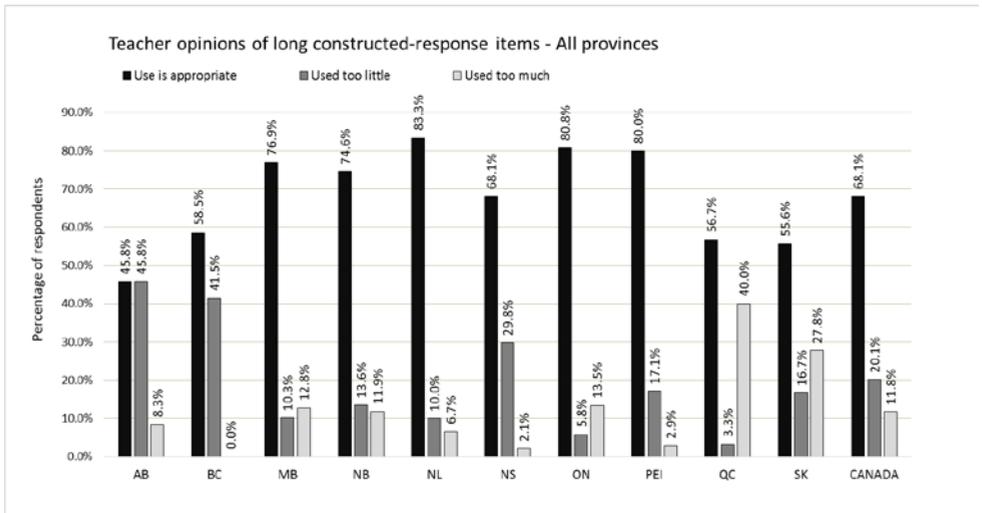


Figure 4.19: National data on teachers' opinions of LSA item types

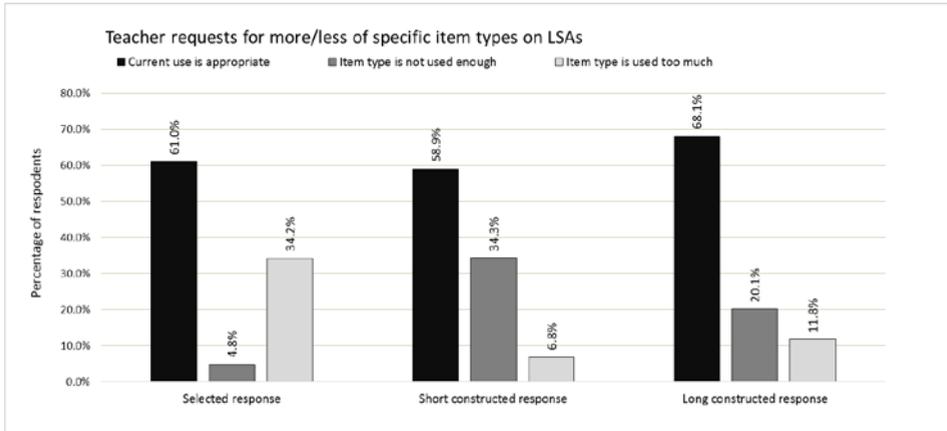


Figure 4.20: Methods by which the results were returned to teachers. Seven responses (including 'other') appeared on the survey, but four (local mark/media, copy only, must ask to see, and not shared) are added here as a result of being commonly written in the 'other' field. Likely these write-ins are under-represented as a result of not being offered as selected choices.

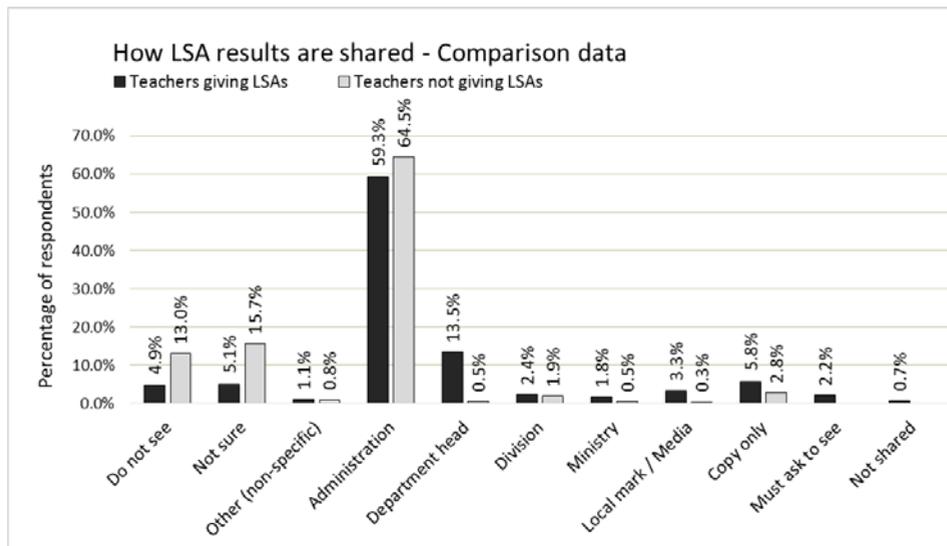


Figure 4.21: National and provincial data on understanding of LSA results. The 'understand' response is self-explanatory, but the other is a grouped response for teachers who do not see, do not understand, or had incomplete understanding of the results data.

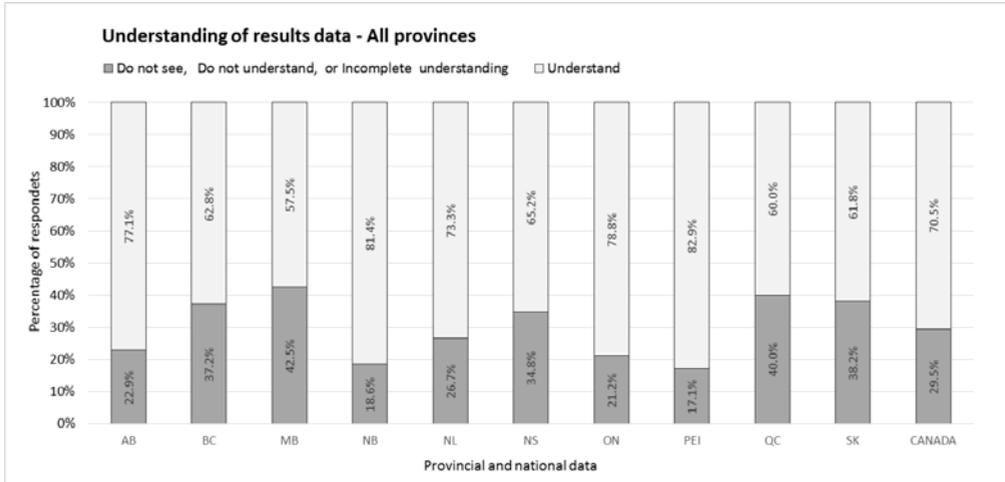


Figure 4.22: Teachers rated their ability to act on results data. The 'can act' and 'some interpretation' responses are as in the survey, but 'can't act' includes several different choices: (a) can't act as a result of poor presentation; (b) can't act as a result of teachers being responsible for analysis; and (c) the write in responses (can't act as a result of poorly timed return of results; results not seen; or questionable data). As above, the write-in responses are likely under-represented in this study. The small proportion of remaining 'other' respondents was not assigned to different categories.

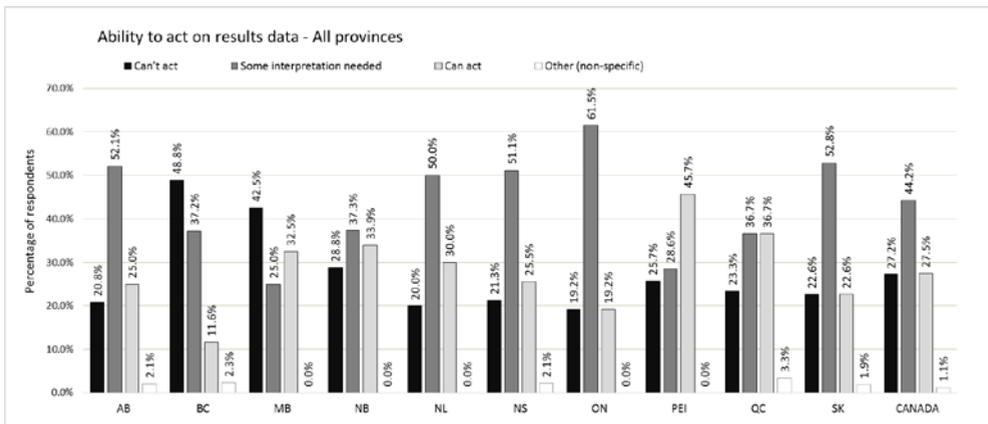


Figure 4.23: Two questions on school accountability were answered by both teachers who do and teachers who do not give LSAs. Responses to the statements were agree (scored as a 1), disagree (-1) or neither agree nor disagree (0).

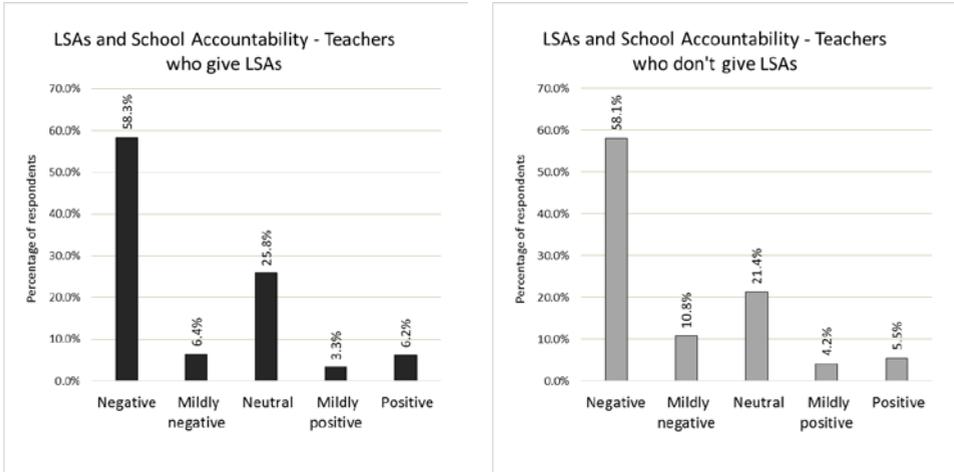


Figure 4.24: Three questions on student accountability were asked with the same response choices as above (4.21; all the questions in this section are of this same format).

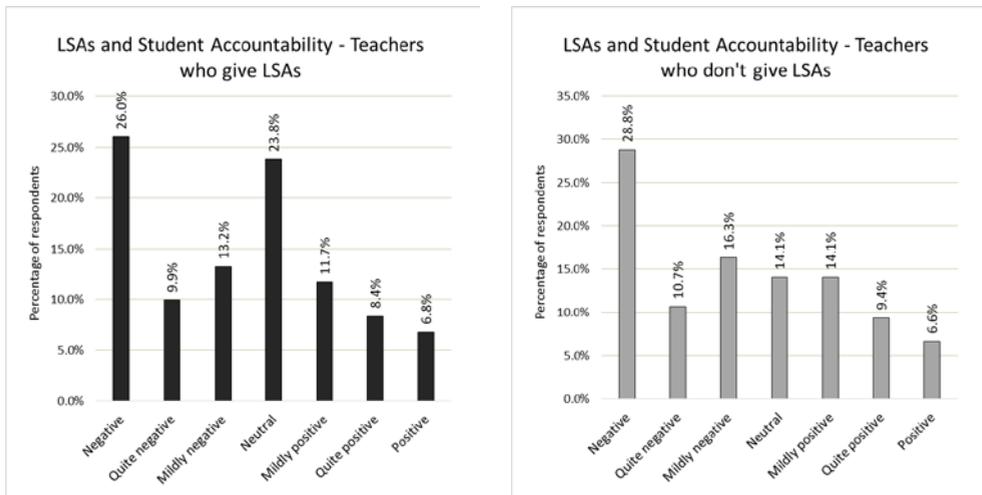


Figure 4.25: Five questions related to the topic of potential for LSAs to assist school improvement were asked. Scores ranged from positive 5 (strongly positive) to negative 5 (strongly negative).

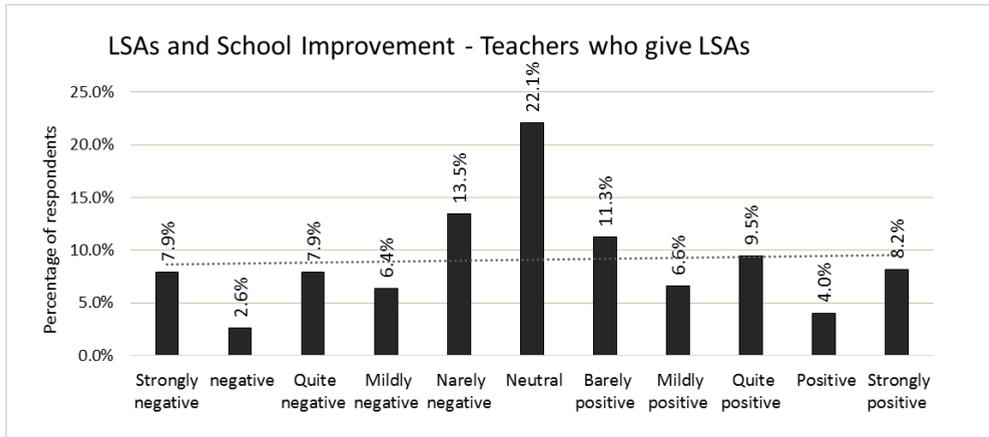


Table 4.26: Distribution analysis for figure 4.23. This distribution is has some extreme opinions apparent in each of the tails, so the variance is high (7.4338), and kurtosis is low (2.4344) indicating a relatively flat curve. The high variance factor also leads to the curve being over-dispersed ($D = 1.4597$).

Using LSAs for school improvement - teachers who give LSAs		Observations: 453
Mean: 0.09271	Standard deviation: 2.7265	Variance: 7.434
Range: -5 to 5	Skewness: -0.02611	Kurtosis: 2.434

Figure 4.27: Five questions related to the topic of potential for LSAs to assist school improvement were asked. Scores ranged from positive 5 (strongly positive) to negative 5 (strongly negative).

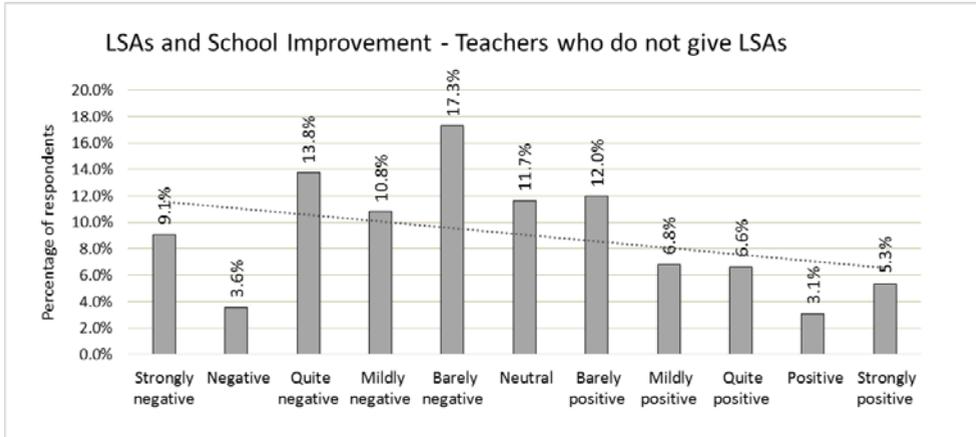


Table 4.28: Distribution analysis for figure 4.25. We see extreme opinions well-represented (large tails, and especially the strong negative opinion), high variance (7.2005), and low kurtosis (2.4001). As in the previous distribution, the curve is over-dispersed ($D = 1.6193$).

Using LSAs for school improvement - teachers not giving LSAs		Observations: 618
Mean: -0.5534	Standard deviation: 2.6833	Variance: 7.201
Range: -5 to 5	Skewness: 0.25122	Kurtosis: 2.400

Figure 4.29: The merged data from both teachers who do and those who do not give LSAs tips towards disagreement with the notion that LSAs can improve schools.

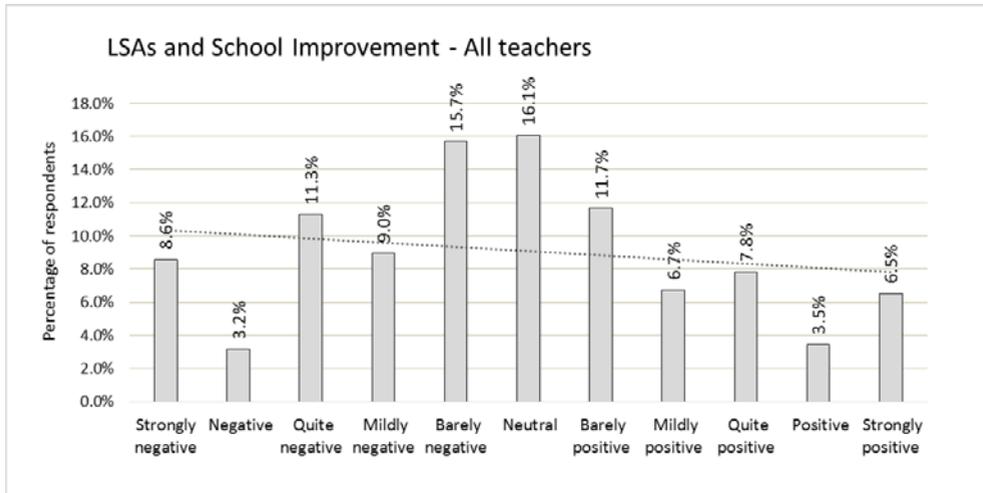


Table 4.30: Distribution analysis for figure 4.27. As with each of the disaggregated curves, this curve is over-dispersed ($D = 1.5666$), has limited skewedness (0.13394), and is relatively flat with a 2.3697 kurtosis factor.

Using LSAs for school improvement - all teachers		Observations: 1071
Mean: -0.2801	Standard deviation: 2.7193	Variance: 7.394
Range -5 to 5	Skewness: 0.13394	Kurtosis: 2.370

Figure 4.31: Five negative statements about tests and their limited utility were proposed, and respondents were asked to agree or disagree. The most notable variations in the curve are high numbers for 'quite positive' and very low numbers for 'negative' (the skewness at 0.1023 bears out a positive trend). Some respondents on the far reaches of the positive and negative scales also means it is slightly over-dispersed ($D = 1.210$).

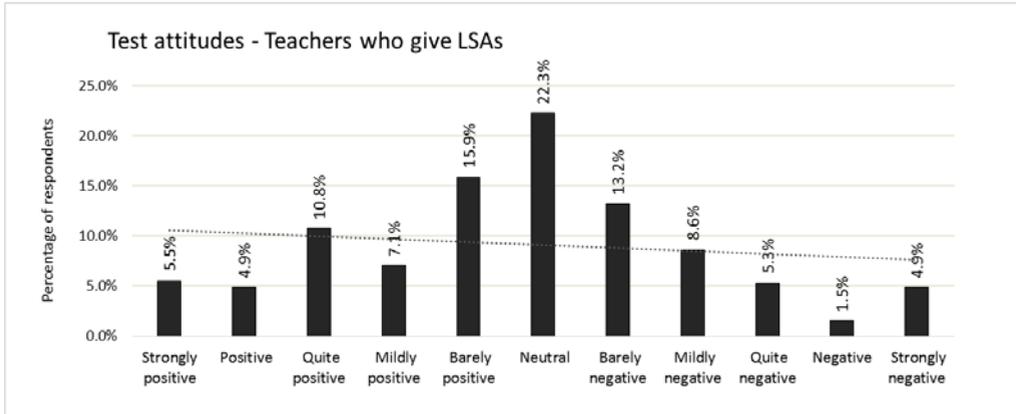


Table 4.32: Distribution analysis for figure 4.29. Some outlier opinions are again reflected in the large variance (5.8930) of the distribution, which is also slightly flat (kurtosis 2.6777). The variance ratio ($D = 1.2610$) reflects some over-dispersal.

Opinion of LSAs - teachers who give LSAs		Observations: 453
Mean: -0.3267	Standard deviation: 2.4276	Variance: 5.893
Range: -5 to 5	Skewness: 0.1023	Kurtosis: 2.754

Figure 4.33: Five negative statements about tests and their limited utility were proposed, and respondents were asked to agree or disagree.

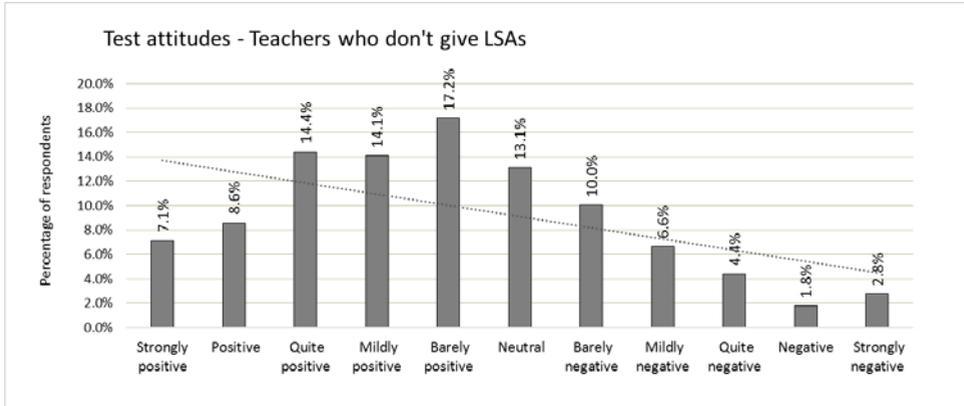


Table 4.34: Distribution analysis for figure 4.31. This distribution is flat (kurtosis 2.6777), skewed slightly left and over-dispersed ($D=1.4794$). The strong positive responses all the way out to the edge of the distribution is the reason for these distortions.

Opinion of LSAs - teachers who do not give LSAs		Observations: 618
Mean: -1.0113	Standard deviation: 2.4292	Variance: 5.901
Range: -5 to 5	Skewness: 0.4002	Kurtosis: 2.678

Figure 4.35: The merged data from both teachers who do and those who do not give LSAs regarding the usefulness of LSAs tends towards agreement with the use of LSA data.

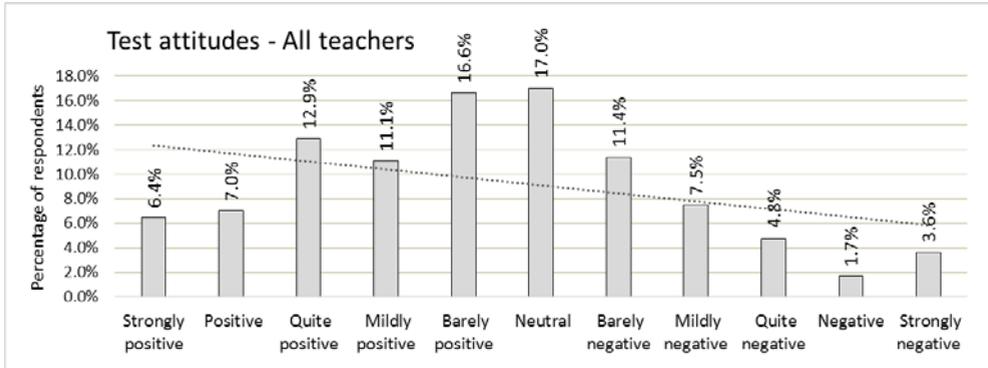


Table 4.36: Distribution analysis for figure 4.33. This curve has a mean to the negative (-0.7218) and is over-dispersed ($D = 1.4040$). Kurtosis and skewness are similar to previous curves, this distribution being flat (kurtosis = 2.6410) and only slightly skewed (0.2670).

Opinion of LSAs - all teachers		Observations: 1071
Mean: -0.7218	Standard deviation: 2.4508	Variance: 6.007
Range -5 to 5	Skewness: 0.267	Kurtosis: 2.641

Figure 4.37: Eight possible uses of LSA data were proposed, and respondents asked to rate these as 'appropriate' or 'inappropriate' uses for the data.

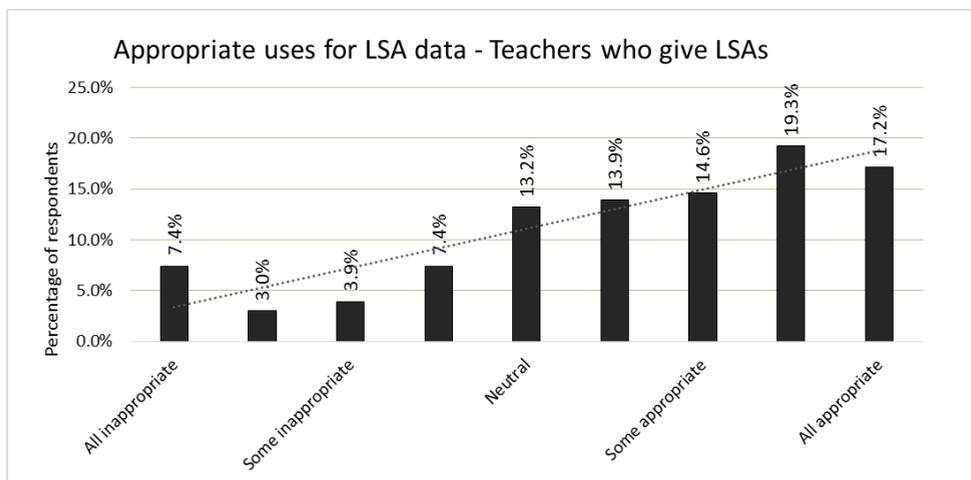


Table 4.38: Distribution analysis for figure 4.31. The distribution is flat (kurtosis 2.6028), has a positive mean (2.3457), has a high variance (22.4546), and is skewed to the right (-0.7155). The curve is also quite over-dispersed ($D = 2.1704$) as a consequence of the grouped responses at or near the 'all appropriate' selection.

Number of appropriate data uses - teachers giving LSAs		Observations: 431
Mean: 2.3457	Standard deviation: 4.7386	Variance: 22.45
Range -8 to 8	Skewness: -0.7155	Kurtosis: 2.603

Figure 4.39: Eight possible uses of LSA data were proposed, and respondents asked to rate these as 'appropriate' or 'inappropriate' uses for the data.

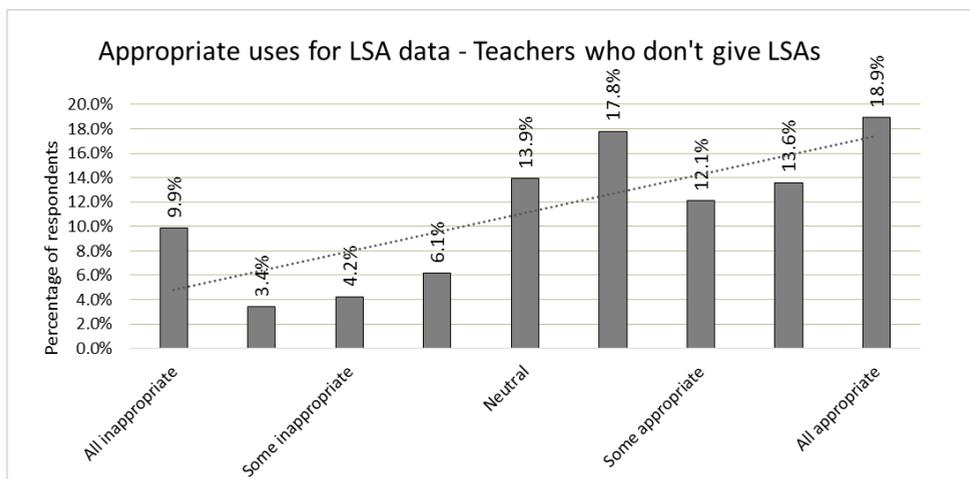


Table 4.40: Distribution analysis for figure 4.33. This distribution is somewhat flat (kurtosis 2.4270), has a positive mean (1.9434), a high variance (24.6370), and is skewed to the left (-0.6198). This curve is also over-dispersed ($D = 2.4778$) as a result of the strongly 'appropriate' trend in responses.

Number of appropriate data uses - teachers not giving LSAs		Observations: 618
Mean: 1.9434	Standard deviation: 4.9636	Variance: 24.64
Range -8 to 8	Skewness: -0.6198	Kurtosis: 2.427

Figure 4.41: Residual analysis for tests and data negative reactivity regressions

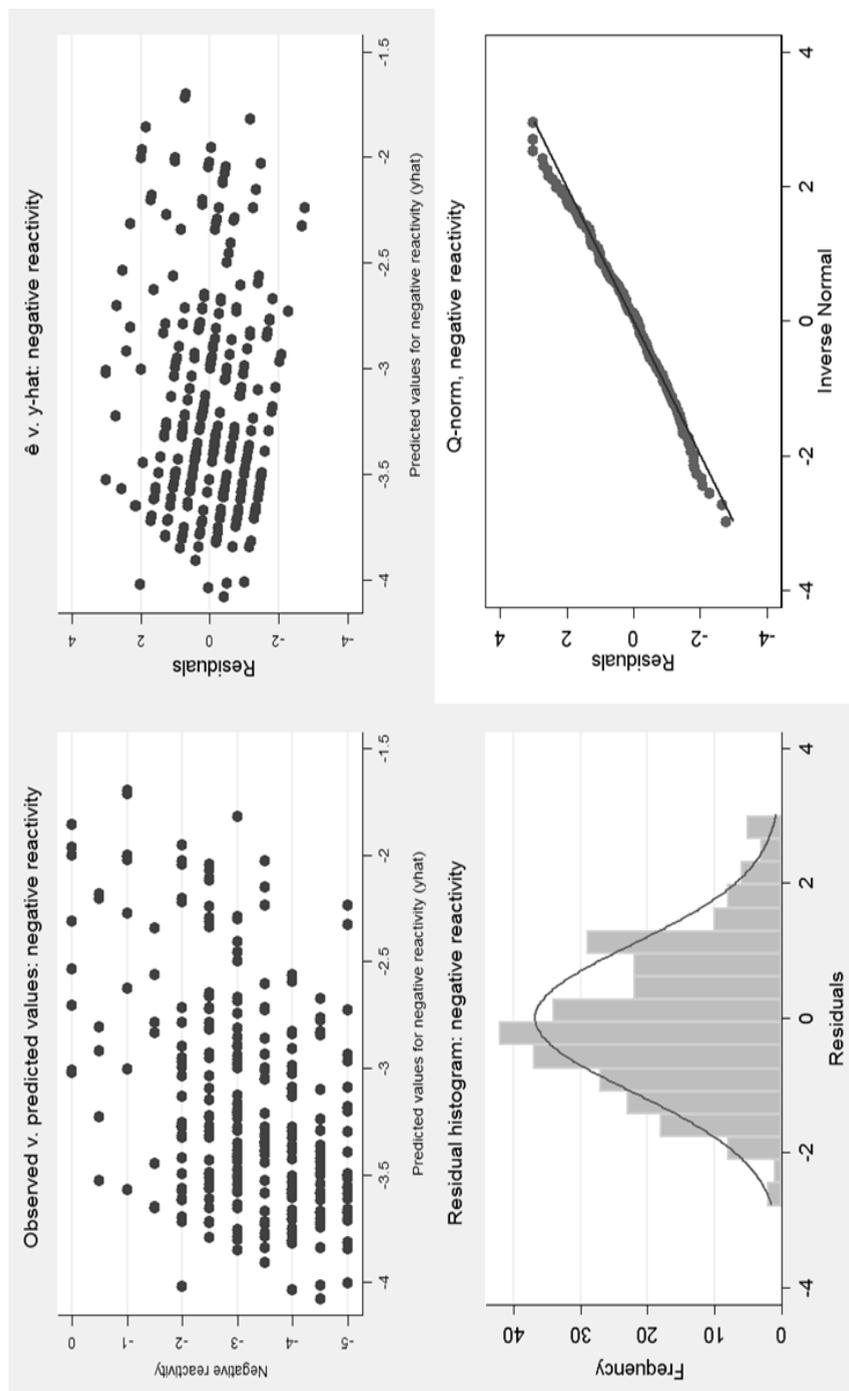


Figure 4.42: Residual analysis for tests and data total reactivity regressions

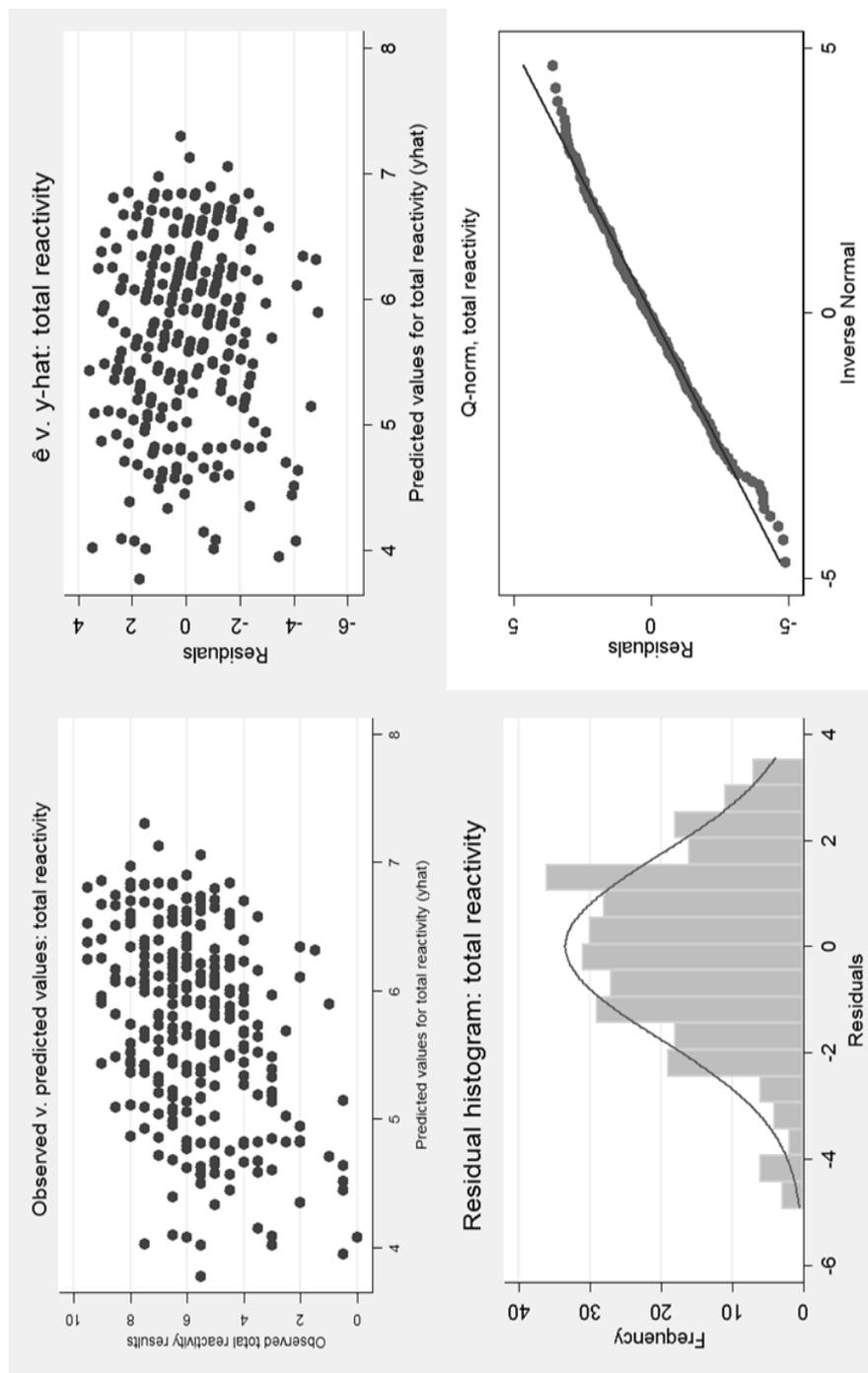


Figure 4.43: Residual analysis for test attitudes positive reactivity regressions

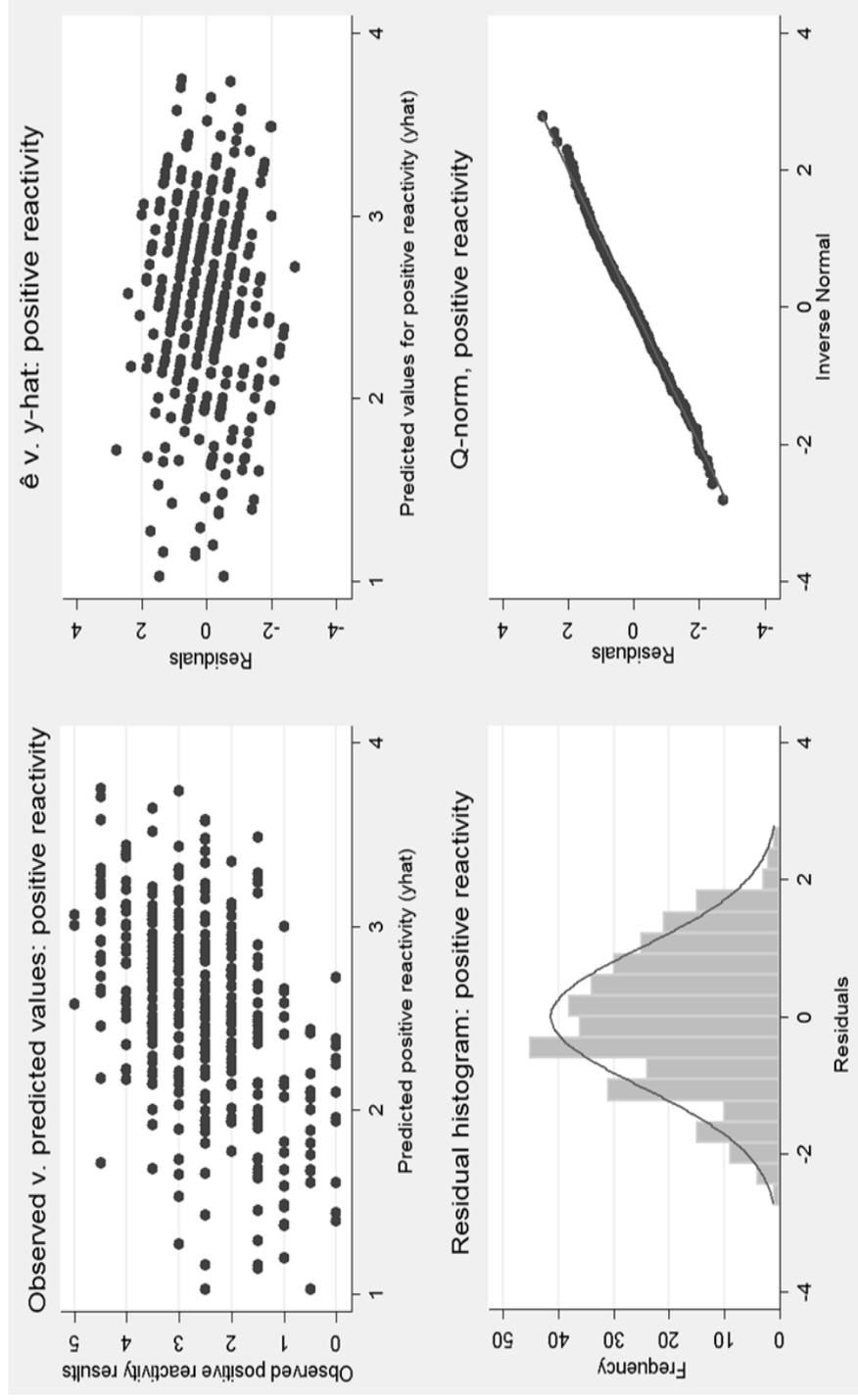
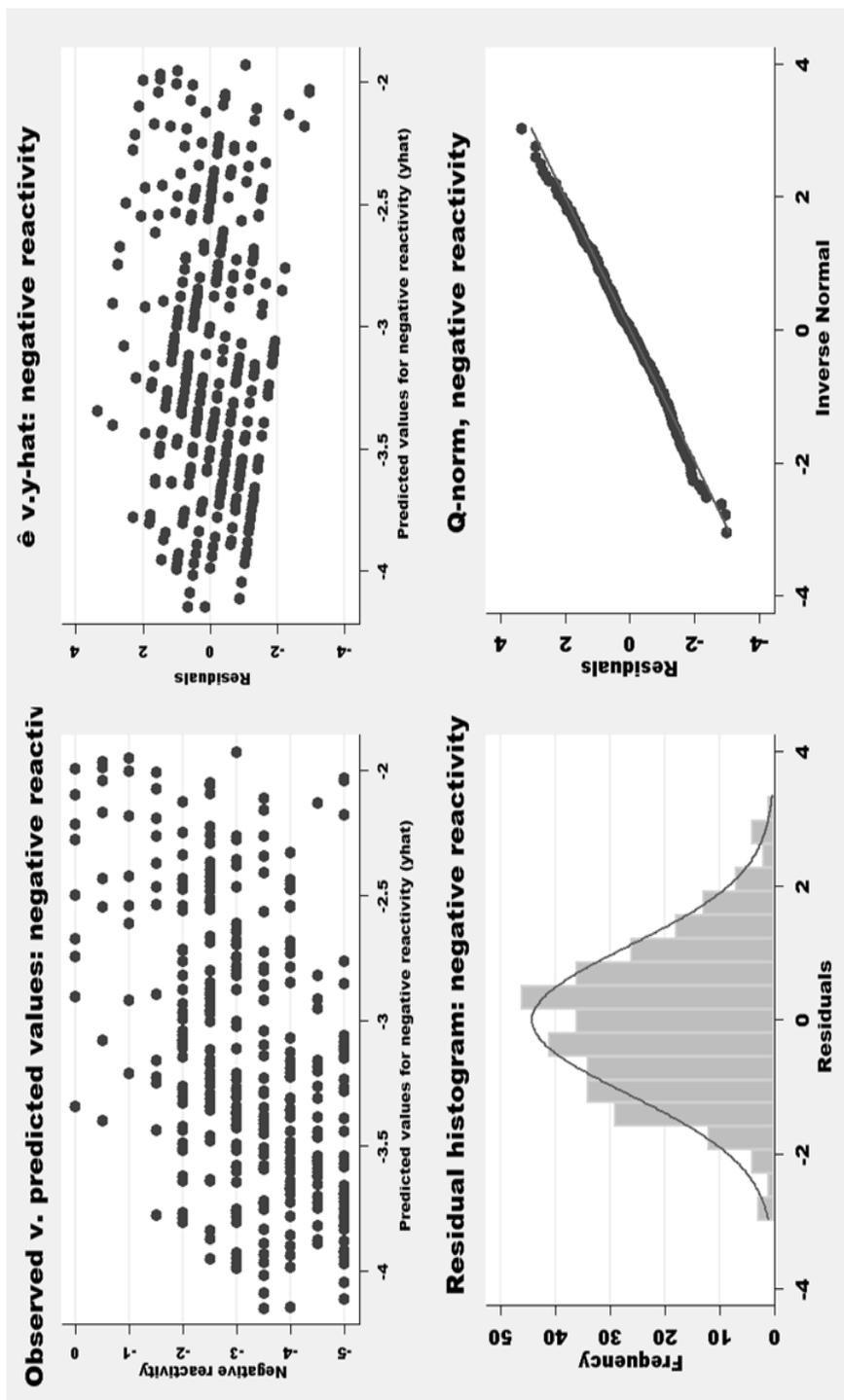


Figure 4.44: Residual analysis for test attitudes negative reactivity regressions



Supports for teachers to use LSA data

5.1 Introduction

Student assessment is a topic given less than adequate coverage in most teacher pre-service programs (Volante, 2006; Young & Kim, 2011; Ungerleider, 2003). This is out of necessity, likely, since to cover assessment in depth would be to exclude other important aspects of teaching. As a result, the support for in-service activities of teachers related to testing and data analysis falls upon the schools, the school divisions and the ministries. They must adequately prepare professional staff for the tasks with which they are charged by provincial LSA programs. The policy choices in Canada related to providing supports for teachers (and more specifically those supports related to assessment, data analysis and data use) will be examined in light of their effects on teachers using the results data to improve instruction.

This chapter is laid out in the following way: (a) a literature review discussing supports for teachers and specifically those related to LSAs; (b) a discussion of the findings from the researcher's survey of teachers all across Canada; and (c) conclusions are presented. These results will address several important independent variables: IV12 - the sharing of data; IV13 - school supports; IV14 - division supports; and IV15 - ministry supports. The three jurisdictional levels will be examined in terms of the number and relative helpfulness of supports provided.

5.2 Literature review

Schools benefit most from LSA initiatives with the engagement of professional teams and improved organizational relationships. The term 'human capital' refers to an individual teacher's skills, knowledge, and abilities but when it is an organizational quality it is 'social capital': the relationships among teachers and those with administration (Fullan, 2011). The school needs a high level of internal capacity to deal with all the data that are produced if they are to use them well (Shepard, 2010). Here organizational theory comes into play affecting school leadership, time use, professional development, and knowledge (Shepard, Davidson & Bowman, 2011). The use of relevant data depends on a culture of collaboration and strong administrative support (Blanc, Christman, Liu, Mitchell, Travers, & Buckley, 2010; Louis, Febey & Schroeder, 2005). In this pursuit, Halverson (2010) advocates a school-wide formative feedback system that would

use assessment results in a formal way with high expectations for improvements.⁷⁸ Focused conversations build internal capacity to interpret and use LSA data and can provide a rationale to reduce prescriptive centralized control of schools, releasing power to local education professionals (Luke, 2011). Whereas testing is a fine tool, it is the people who employ it that really affect outcomes:

Stringent accountability measures, strong curricular guidance, and periodic assessments are no substitute for skilled and knowledgeable practitioners working together in instructional communities to use data to improve instruction. Investments in human capital cannot be bypassed. Data can make problems more visible, but only people can solve them. (Blanc, Christman, Liu, Mitchell, Travers, & Buckley, 2010, p.222)

Corcoran and Goertz explicitly identify explicitly the key elements of school organizational capacity as:

. . . the intellectual ability, knowledge, and skills of teachers and other staff; the quality and quantity of resources available for teaching, including staffing levels, instructional time, and class sizes; and the social organization of instruction or instructional culture. (Corcoran & Goertz , 1995, p.27)

School-level decisions on hiring, budgeting, and leadership are therefore integral to building and maintaining this capacity.

Two related factors may be pivotal in determining the level of organizational support assessment initiatives get: the provision of necessary and effective professional development (PD) related to LSA, and the skills in data analysis that this type of PD should engender. To “raise the level of assessment literacy” in the system some input of time and money must be employed on developing these skills (Morris, 2011, Volante, 2006). The opposite side of the coin

⁷⁸ "A school-level formative feedback system extends the insights from the classroom to the school as a learning organization. A formative feedback system model that would provide useful information about teaching and learning in schools would (a) generate information signals that measure how students performed in terms of an intervention , (b) develop sensor and processor functions to assess information signals, and (c) identify controllers that could actuate this new knowledge in order to adjust the instructional process . The three functions of intervention, assessment and actuation compose the core elements of informative feedback system model." (Halverson, 2010, p.132)

is that if teachers do not know how to interpret these data, it is unlikely that they will use them (Ungerleider, 2006). Even the best intentions to use LSA data are undone if the skills necessary are not taught to educators.⁷⁹ Part of the responsibility for this may lie with teacher training programs (Ungerleider, 2003), but systemic changes require on-the-ground support for staff in every school, and in every jurisdiction (Fullan, 2011). Teachers want and expect useful PD that will give them the skills to deal with classroom assessment issues, and they often consider the more specific, test-related sessions the most useful (Scott, Webber, Aitken & Lupart, 2011). Studies also indicate that those jurisdictions which demonstrate the inclination and investment for developing assessment literacy see the most overall change in teacher performance and adoption of new models of instruction (Schorr, Firestone & Monfils, 2003).

To facilitate the use of these data, there can be: (a) emphasis on the collaborative efforts to work with data; (b) the empowerment of professionals to use sound judgement; and (c) detailed analysis done at the provincial level (Ungerleider, 2006; Young, 2006; Berry, Wade & Trantham, 2009). School divisions are, as a result of their intermediary position between schools and the ministries, in a unique position to make some sense of LSA expectations and data before they are handed down to staff. A pro-active approach is called for at the divisional and ministerial levels to upgrade teacher skills (Volante, Cherubini & Drake, 2008). It should be noted, though, if there is a lack of a coherent assessment strategy (developed together and presented without dissonance) coming from the three oversight bodies (school-based administration, divisional administration, and ministry officials), one can expect little impact in the classrooms (Volante, et al., 2008). In situations that teachers feel least able to meet high standards set by ministries and assessment skills are not developed, the easiest road to improved scores seems to be gaming the LSA systems and preventing least-able students from writing tests to prevent them from taking down average scores (Darling-Hammond & Rustique-Forrester, 2005).

Less consistency in assessment practices might be the result of the school division providing a school organization autonomy or control, but poor results often lead to more invasive oversight and worse. Low-performing schools often have less sharing between staff, or it is episodic in nature (Louis et al., 2005). There is also a turbulent atmosphere in these schools, especially in cases where accountability measures demand sanctions. This means high staff and administration turnover, a revolving door of policies and consultants, and the

⁷⁹ " . . . even if teachers are motivated to respond to testing or other policies that require changes in instructional practice, they may be hard-pressed to actually do so if they do not know what to do or how to do it." (Schorr, Firestone & Monfils, 2003, p.377)

unrelenting eye of district oversight that turns up the heat to unbearable levels⁸⁰ (Mintrop et al., 2009; Willms, 2000). Ravitch claims that tests, regardless of their validity or reliability, can be badly used:

Tests can be designed and used well or badly. The problem was the misuse of testing for high-stakes purposes, the belief that tests could identify with certainty which students should be held back, which teachers and principals should be fired or rewarded, and which schools should be closed - and the idea that these changes would inevitably produce better education. (Ravitch, 2010, p.150)

To avoid the negative outcomes for educators that come with high stakes testing and strict oversight, there has to be 'buy-in' at an early stage (Morris, 2011). High stakes tests can have a demoralizing effect on teachers when they cannot implement instructional change based upon them and thus do not see the value of the results (Crocco & Costigan, 2007). Another side effect of high stakes testing is the de-professionalization of education, and the apparent lack of trust that translates into *lower* expectations for teachers (Dorn, 1998). The willingness of teachers to accept, contribute to, and participate in change is the main element determining the success of any school-based initiative (Louis, Febey & Schroeder, 2005). Whether they do so or not is often based on divisional policies in conjunction with those from the province and the school. A policy that is implemented despite contradictions with those of other jurisdictions has limited chances of inspiring reform (Ryan & Joong, 2005).

Figure 5.1: Summary of supports literature

Topic	Author(s)	Summary statement
Organizational strength	Blanc, Christman, Liu, Mitchell, Travers, & Buckley, 2010	Examines leadership and organizational strength in increasing LSA scores as well as the need to build individual and group data skills.

⁸⁰ "Unless there is substantial improvement in the atmosphere surrounding system-wide testing regimes, teachers are likely to remain suspicious about them. . . Accountability regimes in jurisdictions with well educated teachers such as those found in Canada must be predicated on enabling teachers rather than controlling or 'fixing' them." (Ungerleider, 2006, p.879)

	Corcoran & Goertz, 1995	Reform policy and implementation are examined across three tiers of government. Conclusion that organizational capacity and professional development are key aspects.
	Fullan, 2011	Whole system reform is based on four drivers and neither assessment nor accountability are amongst them. The key, according to the author, is organizational strength.
	Halverson, 2010	Focus on formative feedback and the data related to assessment with the perspective of organizational strength.
	Louis et al., 2005	Looks at how teachers make sense of testing and the effects of this understanding on their practice.
	Pomplun, 1997	Kansas based study which examines the 'path model' basis of instructional change. Concludes that when district, school and teachers <i>all</i> change their practices, the effect is greatest on instruction.
	Runté, 1998	An Alberta example on the issue of de-professionalization and 'proletarianisation' of tasks. Concludes that narrowing of curriculum and lower morale result from LSA policy.
	Shepard, 2010	This paper takes a perspective on Halverson's formative feedback model as well as negative effects such as teaching to the test and high stakes. Important factors appear to be staff readiness to analyze data and organizational strength.
	Shepard, Davidson & Bowman, 2011	Examines the data-driven model using teacher interviews citing specific cases of using or not using data.
Professional development	Borko, Elliot & Uchiyama, 2002	This paper is very positive on the effects of professional development, organizational capacity, and implementation.

	Boyle, Lampranou & Boyle, 2005	Report of the second year of an ongoing study looks at the impacts of professional development for teachers. Different types of professional development are compared as well as different types of impacts.
	Corcoran, Fuhrman & Belcher, 2001	How three school districts had conflicting methods and philosophies about using data and providing professional development
	Desimone, Porter, Garet, Yoon & Birma, 2002	National cross-sectional study on what kind of professional development makes for change in practice. Promotes the use of the 'reformed' type that is active and practical, as well as long-term rather than one-time.
	Parsons & Beauchamps, n.d.	A model for professional development and reform, not specific to LSA testing, but does look at it. AISI is a site-based and multi-jurisdictional model in Canada.
	Schorr, Firestone & Monfils, 2003	A New Jersey study looking at test design, PD opportunities, and stakes. In many cases, they find, teachers take on strategies in name only, the data have limited reactivity, and PD is test-based and ineffective.
	Scott, Webber, Aitken & Lupart, 2011	Looks at Alberta work and finds that professional development important, as is engagement. The ties of curriculum to test and expertise are also important in changing practice.
	Sharkey & Murnane, 2006	Examines formative assessment choices in light of assessment purposes, weaknesses, and teacher professional development.
	Ungerleider, 2003	Examines testing policy from the perspective of teacher engagement, common standards, common misuses of the data, and the need for PD.
	Ungerleider, 2006	This paper covers topics such as teacher involvement, cross- purpose testing, results reporting, and the non-use of data.

	Volante, 2006	Starting with a look at provincial test history, the author lists ten criticisms, ten needed supports (including professional development) as well as a focus on technical quality of tests.
Facilitating data use	Anderson, Leithwood & Strauss, 2010	Examines organizational factors in data use across jurisdictions. This study has a similar design model to the author's but a different theoretical model.
	Armstrong & Anthes, 2001	Study of six effective data-use districts that draws general conclusions about school climate and data-use structures.
	Berry, Wade & Trantham, 2009	Paper concludes that school working conditions are related to both empowerment and productivity.
	Coburn & Turner, 2011	Data use is a construct with a complex web of factors affecting it, thus: jurisdictions, leaders, tools, routines, PD, stakes, etc. are examined.
	Darling-Hammond & Rustique-Forrester, 2005	Examines the effects of high stakes in Connecticut, Kentucky, Vermont cases. Authors conclude that policy design and implementation are key, PD is needed and that buy-in can prevent negative reactivity effects.
	Earl & Fullan, 2003	Compares Ontario, Manitoba and United Kingdom accountability systems and the reactions to it. There is some examination of policy implementation and also on data analysis.
	Halverson & Thomas, 2007	Paper argues that resources teachers (SSTs) are in-house data experts using 2 case studies related to applying this to instruction at school level.
	Kerr, Marsh, Ikemoto, Darilek & Barney, 2006	Paper comparing data use in three US districts. Concludes that the most preferred are examples of multiple measures.
	Lachat & Smith, 2009	This paper looks at three urban schools' experiences trying to use LSA data. School leaders are important as are also data

		analysis skills and PD.
	Lee & Wiliam, 2005	A case study of two teachers and their changes in practice in formative assessment. This is very informative on the factors that made a change in practice possible, attractive, and sustainable.
	Schildkamp & Kuiper, 2010	Study examines how are data being used and what variables hinder or help in the Netherlands. Six sites were selected and qualitative methods used.
	Supovitz, 2013	A New Jersey district-based study on how teachers engage with data to inform instruction. An interesting idea presented here is that it is not data analysis skills that facilitate data use but an ability to construct meaning from the data.
	Wayman, Cho, Jimerson & Spikes, 2012	This paper looks at several ways LSA data are used and includes the survey questions about their attitudes. These were adapted for use in this study.
	Weinbaum, 2009	Delaware-based 10 state study of AfL enhancement at high schools. The paper focus on high school teams' implementation of AfL vision which shows success only from cases with strong state or school leaders. The rest wallowed and grew little This is basically the story of overworked, under-supported early adopters who didn't change practice as a result of constraints.
	Willms, 2000	Examines how testing is done, and how it can be better. Topics include test design, reactivity effects, jurisdictional roles, and teacher preparedness to use data.
	Young, 2006	Focus on leadership and other means to get data used namely alignment to curriculum and practices to facilitate data use.
	Young & Kim, 2011	A comprehensive literature review on the uses of data. Includes lots of good detailed information and a stacked bibliography.

Staff support	Crocco & Costigan, 2007	Interviews with New York City new teachers examining autonomy, high stakes, curriculum narrowing, and prescriptive test administration.
	Debard & Kubow, 2002	Paper on implementation issues, and the lack of 'bottom up' being a stumbling block of LSA mandates (policymakers need to consult).
	Dorn, 1998	Paper focused on high stakes accountability and its political consequences which advocates reframing the topic of educational reform.
	Falk & Ort, 1998	Looks at teachers' involvement in scoring and how it improves confidence, understanding, and classroom testing practices.
	Hargreaves, Crocker, Davis, McEwen, Sahlberg, Shirley, & Sumara, 2009	A critical analysis of AISI program which includes several papers, all of which see local autonomy as the key to success of this Alberta initiative.
	Louis, Febey & Schroeder, 2005	Paper looks at how teachers make sense of testing and the effects on practice from a teacher perspective.
	Wideman, 2002	A study of 25 teachers in two school boards, this is action research with LSA data looking at the changes year-on-year. The focus is the professional development model, not using results.

5.3 Preliminary hypothesis

H 5-1: Teachers who feel supported in their professional practices will be more inclined to adopt data-informed instructional techniques. Thus, positive opinions about provided supports and the recognition that supports are available will have an affirmative impact on **total reactivity** scores and more specifically increase **positive reactivity**.

5.4 Results from surveys

The survey results will be presented by province and nationally looking at the key independent variables in this chapter. Most of the statistical analyses are left to the end, and will look at correlations between these independent variables and the dependent variable. Finally, trends will be identified and conclusions drawn. These are the variables discussed in this chapter:

Supports results: These are provincial data which test hypothesis **H 5-1**

- IV12 (independent variable 12) - sharing of data
- IV13 - school supports
- IV14 - division supports
- and, IV15 - ministry supports

Survey respondents were asked a series of questions related to the supports they receive from their schools, school divisions, and the provincial ministry. They were asked rate how many supports to which they had access as well as how helpful they found them. Each support provided was considered a positive, but the rated helpfulness could be positive (helpful or very helpful) or negative (not helpful or not provided). Thus survey responses converted from ordinal choices (often based on Likert scales) into cardinal values based on the judgment of the researcher. These values were combined to create the values used in the regressions.

Note that these variables are seen individually and analyzed further in the chapter-ending charts and tables section (section 5.8), and that the values given to survey responses for regression purposes can be seen in **Annex 2**.

5.4.1 National results

- IV12 - sharing of data (see **Figures 5.7 and 5.8**)

The frequency of data-sharing is mixed across the national sample. While the most common response was 'sometimes' (45%), the never response was more common than 'always' (31% to 24%). This indicates there are relatively few jurisdictions where there is either mandatory compliance or communities of teacher learning which take it upon themselves to do this.

- IV13 - school supports (see **Figures 5.9 – 5.12, and 5.15**)

Supports provided at the school-level were the most common across Canada, and led responses in four of the six supports categories. Overall, 51% of the supports received by respondents came from the school-level. These supports were considered helpful most commonly (49% of responses) followed by very helpful

(27%). Less appreciate voices responded they were not helpful or not provided in smaller proportions.

- IV14 - division supports (see **Figures 5.9 – 5.11, 5.13 and 5.14**)

Division-level supports were the second most commonly provided across the nation making up 34% of all received supports. Divisions led in the provision of professional development, likely because larger numbers of attendees at these sessions allows economies of scale as well as the ability to share a common message to all staff. The numbers for considered helpfulness of supports fall quite substantially at the division-level compared to the school-level. Yet 48% of respondents consider these helpful, but only 10% rate them as very helpful. More to the point, 18% rated them as not helpful and 25% reported they were not provided.

- IV15 - ministry supports (see **Figures 5.9 - 5.11, 5.14 and 5.15**)

Ministry-level supports are the least commonly reported by respondents to the surveys making up 15% of total provided supports. The ministry has less than 10% penetration in three of the six supports categories but leads in online/printed guides, has 14% coverage in professional development, and 11% for assessment teams. The helpfulness ratings for these supports are quite low - 37% of respondents find them helpful, but 36% report they are not provided. And while 7% find them very helpful, but 20% do not find them helpful at all.

- Summary of supports data

The data regarding supports seen from the national perspective paints a positive picture of school-level supports (in that they are commonly available and are considered helpful), but less positive perspectives on divisional and ministry supports are apparent. Some supports are much easier to deliver at the school level. Professional learning communities (PLCs) for example, need to meet quite regularly to do the work they are expected to do (Dufour, DuFour, Eaker & Many, 2006). Administrative support also is 'in-house' at a school, whereas a visit by a superintendent or director would not be as common.

Schools do not lead in the provision of Professional Development supports since it regularly has a substantial dollar value attached when a keynote speaker is booked for the event. Yet not all PD has to follow this model, and in-school PD sessions can have more directly relevant outcomes than centrally planned events (Nagy, 2000; Schorr, Firestone & Monfils, 2003). Divisions lead in providing PD as well as large portion of coaching support, and this follows from the sheer scale and budget of the division, the expertise of the centralized coaching staff, and the allocated budget used to support all teachers in this way. These supports are, by and large, found to be helpful, although they are not as well-rated as those from the school level.

Ministries are almost absent from the supports discussion, and provide less than one in six of the supports reported. They lead only in the provision of

printed and online guides (which only makes sense since ministries write both tests and the testing procedures) and have semi-respectable numbers for PD and assessment teams. The ratings for helpfulness were the most divergent here as ministry supports were poorly rated or not available to 56% of respondents.

The provincial workshops were very helpful but rarely provided.

- **Anonymous survey comment**

The FILA provincial assessments need more support as nothing is done with them or about them. Sometimes, we need to ask for the results! Our results are very poor. The students and teachers are frustrated about them. We are asking for support, yet nothing....and the cycle continues!

- **Anonymous survey comment**

5.4.2 Provincial results

Alberta

- IV12 - sharing of data (see **Figures 5.7 and 5.8**)

There is a greater prevalence of data being shared in Alberta than was true nationally. Data were 'never' shared 10% less commonly, and 'always' shared 10% more frequently.

- IV13 - school supports (see **Figures 5.9 – 5.12, and 5.15**)

Alberta's level of supports from schools (as compared to divisional or ministry supports) is 7% higher than national figures show (58% to 51%). They are also considered helpful by 89% of teachers (the national average is 76%). The model of providing supports is much like many other provinces, but the high helpfulness ratings differ substantially.

- IV14 - division supports (see **Figures 5.9 – 5.11, 5.13 and 5.14**)

There are fewer division level supports reported in Alberta than is true nationally (27% to 34%). They are also considered less helpful by respondents here. In Alberta 37% find them either helpful or very helpful, whereas the national average for this group is 58%. The difference is made up with higher figures in the 'not helpful' (14% higher) and 'not provided' (7% higher) groups.

- IV15 - ministry supports (see **Figures 5.9 - 5.11, 5.14 and 5.15**)

The number of ministry supports in Alberta is very close to the national average. All helpfulness categories are marginally above national figures with that proportion (6%) all coming from the consequentially lower 'not provided' numbers.

- Summary of supports data

The data from Alberta show that most teachers in Alberta have access to a sufficient number of supports and that they find these supports generally helpful. Those are certainly positive indicators. Being responsive to data often depends on having access to professional development or professional learning communities to help individual teachers manage the analysis and interpretation needed to examine the data, but then also using provided supports to devise strategies to work with the data in constructive ways.

Well, we can request from the province that a visit by one of the exam managers or someone associated with the exam process to come out and help us with the data. And there is some drilling down that you can do if you look at the statistics of it where you can make some broad-based changes to your delivery practice based on what you see. . . That's the kind of thing you can request and they are more than happy to do that. I'm not sure a lot of teachers take advantage of that as much as they should, but it is a possibility.

– **AB, High school Science teacher, male**

Respondents from Alberta also indicate, as is true nationally, that school-based supports are the most helpful and most common. There is no doubt more opportunity for cost-effective PD at the school level, but the quality is likely greatly variable. School's discretionary budgets may restrict the types of opportunities available at this level, whereas professional learning communities (PLCs) depend more on in-house expertise (itself a function of the size of the staff and the qualifications of the teachers).

Our principal was kind enough to give us [substitute teacher] time to analyze the results and determine if there were areas that as a grade team we felt we could improve on . . . and that was helpful.

- **AB, Elementary English teacher, female**

In our department, in my school we work together collaboratively which I guess you could call a PLC, but it doesn't hit a lot of the defined features of PLC. We as a department will seek out PD or time to work together on improving instruction but rarely will there be any outside sort of imposed or offered set of supports for that work. - **AB, High school Math teacher, female**

In all, Alberta's level of supports seems to be in line with their generally high levels of positive reactivity. It is the second most positively reactive province

(see Chapter 3), and supports are hypothesized to be a telling predictor of increased positive reactivity effects.

British Columbia

- IV12 - sharing of data

British Columbia teachers were much less likely to report data being shared between staff than any other province. The 3% rate for data 'always' being shared is 11% less than the second lowest rating (Manitoba). They also have by far and away the highest rating for data never being shared (57%) which compares unfavourably with the national average of 31%.

- IV13 - school supports

Half of the total number of provided supports come from the school level which is very near the national average (51%). The reported helpfulness of supports paints a different picture. Only 34% of respondents indicate that supports were either helpful or very helpful. Fully 60% said they were not provided (note that this does not contradict the 50% of provided supports finding since this is a proportion of only those supports that *are* provided). Like Alberta, the jurisdictional provision of supports is much like the other provinces, but the helpfulness ratings are in this case vastly different.

- IV14 - division supports

As a proportion of the total number of supports provided, division-level supports are under-reported in British Columbia (29% to 34% nationally). Only 29% were considered helpful or very helpful, 7% were not helpful, and 65% of respondents indicate they are not provided.

- IV15 - ministry supports

The ministry provide 6% more supports as a proportion of the total in British Columbia but their ratings for helpfulness remain abysmal – 23% report helpful or very helpful supports, 13% find them not helpful, and 63% indicate they are not provided.

- Summary of supports data

It has been noted already that British Columbia is an outlier in many respects when looking at responses to survey questions. In the variables examined above, this trend is again clear. The rate of 'not provided' responses from BC teachers more than doubles the national averages for school- and division-level supports (it is only 75% higher than the ministry indicator). There can be little doubt that teacher-respondents from this province feel that supports are not available while the perspective from school divisions does not align directly with this sentiment (perhaps as a result of the labour unrest).

There is nothing I've seen directed or our administrator has never said anything to us. It wasn't like a priority for him [the principal] to even let us know, which I totally support.

– **BC, Elementary homeroom teacher, female**

And the province is good, I mean if people are willing to accept it. They do item analysis, they send that out. . . They provide some really good analysis of what comes out of the assessments. It is just that when it is such a political controversy, whether or not people look at that, I don't know.

– **BC, Division staff, male**

Even at the most basic level, teacher interactions to share LSA results data, British Columbia teachers do not come anywhere close to the national average. Just 3% of respondents state that data are always shared between staff. If there is a data-informed educational climate in BC schools, the Foundations Skills Assessment tests (FSAs) are not a part of it.

I do not use any for support with grade 4 FSAs. These tests do not accurately assess students. I do use many of the [listed professional development strategies] to improve both my instruction and assessment practices *within the classroom context* [italics from original], but I don't use the [FSA] assessment results. I look over the results merely to see how my students have done and if their results mesh with what I see in the classroom.

– **Anonymous survey comment**

The Foundations Skills Assessments and diploma exams given in British Columbia schools represent a large investment of time and money. Without the supports needed to analyze and use the data in suitable ways, the ministry is not getting a reasonable return on this investment. It may be aspirational for the ministry to state that these results are “intended as a resource to support instruction” since so very little support has apparently been provided to teachers to make this a reality.⁸¹ The reactivity data from Chapter 3 show that teachers in British Columbia are the least positively reactive and the most net negatively reactive in Canada. Another perspective on these data would be to say that without sufficient and helpful supports, the 'low road' of negative reactivity is much more commonly taken by teachers.

⁸¹ British Columbia Education, retrieved Aug. 9, 2014 from:
http://www.bced.gov.bc.ca/assessment/fsa/pdfs/fsabrochure_print.pdf

Manitoba

- IV12 - sharing of data

Manitoba teachers reported the second highest national rate of never sharing LSA data between teachers (50% of respondents). The national average is 31%. There is 10% less than average sharing in the 'sometimes' category, and 9% less sharing in the 'always' category.

- IV13 - school supports

Manitoba teachers report a much more equitable mix between jurisdictions that provide supports than is true nationally – only 15% separates ministerial and school-based supports (respectively the least and most commonly indicated). School support levels come in at 40% and the helpfulness ratings are not significantly different from national averages (77% find them helpful or very helpful, and 23% find them not helpful or not provided compared to 76% and 24% nationally).

- IV14 - division supports

Division supports make up 2% more of those provided to Manitoba teachers compared to national figures. These are considered much more helpful (by fully 20% over national scores in that category) and are 'not provided' just about as regularly as in Canada as a whole (24% to 25%).

- IV15 - ministry supports

Ministerial supports are 10% more common in Manitoba than is true nationwide. The helpfulness ratings are also higher (66% for 'helpful' and 'very helpful' responses compared to 44% nationally). With more ministerial support reported, the 'not provided' response is 17% lower than national numbers.

- Summary of supports data

Manitoba has a good record of providing supports to teachers related to LSAs according to the respondents from this province. The supports are fairly evenly spread across jurisdictional levels and the helpfulness ratings are all above national norms. This being the case, it bears asking why reactivity scores in Manitoba are in the middle of the pack, and they rate so low for positive effects if these supports are in place and effective. Interview respondents did indicate that PD was more common in years past (when the current LSAs were new) but has tailed off in recent years.

The first like maybe year or two [of implementation] . . . there was sessions and things offered, but I don't know if they are still being offered. I haven't sought them out because I feel I don't really need to anymore. – **MB, Middle years Math teacher, female**

When the assessment first came down there was a lot of panic because the thought of standardized tests is scary to most teachers. . . So as a division we sat down and we actually created a package that had a lot of teaching tools in it, not just assessment pieces, but teaching tools and strategies.

- MB, Elementary school principal, female

And I have done a couple in-services on doing the assessment, and how you can best do it in your classroom. And once that all happened, now it is just practice. People don't even need to talk about it, like they hardly even talk about it anymore.

- MB, Elementary school principal, female

The low data-sharing numbers might be another reason for this disconnect. Used or not, supports and training do serve to clarify the purpose and intent of the LSA program, and they are necessary to connect these dots if the data are to be effectively used. Yet where data-sharing does not occur, there are reasons why:

And it is very unclear as to why they are collecting this data. Like, I have never seen a report generated from this, like, it has never really come up. So I don't know how much value is placed on it and what we're doing it for, because all of the training sessions are very vague. Everyone interprets what we are supposed to be doing for this very, very differently and there is no standard of how we are supposed to be actually assessing the students to see what they know.

- MB, Middle years Math teacher, female

The division itself has a literacy coach and now put on, put a position in as a numeracy coach for particular schools but not the entire division. . . . As an administrator if I want to give them extra time, say, as a teacher, to sit down and look at results and discuss it with colleagues, then I will. That comes out of my time and my planning. It is not done divisionally.

- MB, Elementary school principal, female

One aspect of policy implementation that may warrant more investigation based on these findings is how quickly a policy initiative loses momentum if it does not remain a focus of not only supports, but also incentives which might inspire continued diligence (incentives are the topic of Chapter 6).

New Brunswick

- IV12 - sharing of data (see **Figures 5.7 and 5.8**)

New Brunswick teachers report a high level of infrequent ('sometimes') data sharing and about equal proportions between data being shared 'always' or 'never' (24% and 23% respectively).

- IV13 - school supports (see **Figures 5.9 – 5.12, and 5.15**)

This province follows the supports model common in two other Atlantic provinces (Nova Scotia and Newfoundland and Labrador) where, as a proportion of the total, most supports come from the school level (58% here and 51% nationally). Respondents from this jurisdiction also indicate a high level of satisfaction with school-based supports. 85% stated they are helpful or very helpful (nationally that number is 76%). Low numbers for not helpful and not provided supports were recorded - 4% and 12%, both lower than national numbers).

- IV14 - division supports (see **Figures 5.9 – 5.11, 5.13 and 5.14**)

The proportion of supports coming from the division level is just about the same as national figures (32% to 33%). The difference in this province is that the 'helpful' category was much more highly rated (62% of respondents) than national data (48%). Very helpful numbers were slightly higher than average (14% to 10%), and negative responses were both lower than average ('not helpful' 6% lower, 'not provided' 11% lower).

- IV15 - ministry supports (see **Figures 5.9 - 5.11, 5.14 and 5.15**)

Proportionally fewer supports come from the ministry level in this province since 46% of respondents indicated they are not provided to them at all. They are not considered as helpful as in the national data, but also show a lower proportion of 'not helpful' responses. The issue seems to be more that they are not available than the fact that they are considered inadequate.

- Summary of supports data

Data from New Brunswick are the first to be examined that set a model of support provision followed by Newfoundland and Labrador, Nova Scotia, and also Ontario. They deviate from each other on considered helpfulness somewhat, but the provision model is common (Alberta and British Columbia have similar proportions but differ greatly in all regards for helpfulness, so have not been included in this group). In these provinces the lion's share of supports are coming from the school level. These local supports are also well-regarded.

We have a pretty good support team in literacy. . . At the school level as well, we have teams at our school, so we really work with each other but we also have from the district some literacy leads who come in and help.

- NB, Middle years homeroom teacher, female

Certainly if one teacher is doing something better or getting higher marks on something then we would be looking at that and saying well, 'What are you doing because obviously what you're doing is making a bigger impact.'

- **NB, Middle years homeroom teacher, female**

It develops an opportunity for colleagues to work together so that if teacher A, if their students across the board haven't done well in one particular area and teacher B, their students have been exceptional in that area, then it develops an opportunity for those two educators to have a conversation about what strategies are used in your class, that hopefully create improvements in the other class. Now that is a difficult path to go down but I still think that as educators we are responsible to do the, to provide the very best education we can for our, for all of our students. - **NB, High school principal, male**

What is less clear is the connection between the local-provision model for supports (when related to LSAs and data use from the assessments) and reactivity, specifically positive reactivity. Of the four provinces cited above as similar with supports, three of them are quite high in total reactivity ranking national, and the other (Nova Scotia) is the second least reactive province. The three provinces are also ranked one, two, three in negative reactivity effects. To be fair, New Brunswick and Newfoundland and Labrador also rate first and third (respectively) in terms of positive reactivity effects. Still, even before crunching numbers, the data appear to point away from the preliminary hypothesis that supports provided and considered helpful would increase positive reactivity.

A factor that might be more telling in the case of this group of four, then, is the relative lack of ministerial supports. The closest link across all the independent variables examined in this chapter is between New Brunswick and Newfoundland and Labrador regarding ministerial supports. These two provinces rate very high on both positive and negative reactivity effects. They also report very high proportions of 'not provided' supports from the ministry level.

I really do not get an opportunity to discuss these results in detail. In my experience, very little is given to assist with the assessment results. In most cases, it is up to the individual teacher to seek out assistance of any kind. – **Anonymous survey comment**

To me, I wish we had the flexibility, umm, to have extra time for literacy and extra time for numeracy. I truly believe that if the province is going to continually state that our mandate is to improve

those two key areas, then they have to realize that not every student is on the same timeline. . . - **NB, High school principal, male**

Newfoundland and Labrador

- IV12 - sharing of data

Data sharing in Newfoundland and Labrador is more frequently reported than national scores. The 'always' shared category is 9% higher (33% to 24%) and the difference comes just about exclusively from the 'never' shared category (22% to 31% nationally).

- IV13 - school supports

There are more supports that come from schools in Newfoundland and Labrador by 6% as compared to national numbers. A 79% proportion of respondents find these supports helpful or very helpful (the national average for these groupings is 75%). That said, 7% nationally find these same supports not helpful, whereas that figure is 15% in this province.

- IV14 - division supports

Division supports are just about on par with the national level for this metric, as are both 'helpful' categories. There are 10% more teachers who find division supports 'not helpful' here (as compared to national figures), but 11% fewer who indicate they are not available.

- IV15 - ministry supports

As a proportion of the total, supports from the ministry are reported to be 5% less common than what is true across Canada (10% to 15%). There is also a larger proportion of teachers who indicate that ministry supports are not provided – 41% in Newfoundland and Labrador, while only 36% nationally. This difference is in large part from a lower proportion of respondents in the helpful category (33% down from 37% nationally), yet the other scores are similar to national ones.

- Summary of supports data

When you consider the total amount of reactivity apparent in Newfoundland and Labrador, a link to supports would be the first obvious place to examine to see why this is occurring. Professional development, professional learning communities and other supports that are directed at provincial assessment goals, regardless of their source, should provide teachers with the skills they would need to act on the data provided to them from LSAs. It has been noted in the examination of Manitoba data that this link did not appear to be obvious, and this is also true here in Newfoundland and Labrador data. Supports are more commonly school-based in this province, but ratings of the helpfulness of these supports are all lower than is true nationally.

The school board wants you to do this, and the one time that we do it's we get together and say look this year we're going to emphasize [subject A] or this year we're going to emphasize [subject B]. But then unfortunately we go into our silos and we sit alone and we remember hopefully what was said in September and we execute that. Help from the school board? None. Help from the school itself, the administration, other than photocopying . . . none. So, no, you are on your own. That one meeting in September drives it all. From then on you are on your own.

- NL, High school Science teacher, male

Supports appear to be readily available to teachers in the province. The number of respondents indicating that supports are not provided is 11% lower than national average for both school and divisional supports. The numbers are only slightly higher in Newfoundland and Labrador (by 5%) for unavailable ministry supports. What is somewhat surprising is that the most commonly provided supports are also the ones considered least helpful. The proportion of respondents consider these supports helpful is lower than national figures for school and division level supports, but just about on the national average for ministry supports. It appears that familiarity breeds contempt.

It would be great if we could have the time to focus on that, to get together several times as teachers and brainstorm what we really could do other than go away and do, but plan some things. It would be phenomenal. We're always going 'we would love to increase students' performance and this is the way to do it,' but they run off on all their other little projects. No, zero support.

- NL, High school Science teacher, male

So correlating supports to Newfoundland and Labrador's provincially highest 'total reactivity' rating would seem to be a stretch at this point. It may prove to be true that supports are not the most telling factor overall in this analysis of policy factors that promote reactivity effects.

Nova Scotia

- IV12 - sharing of data

Nova Scotia respondents indicate that sharing results is about as common in this province as nationally, although 8% more 'sometimes' responses and 8% fewer 'always' responses were registered than is true for the national average. The 'never' response, at 30%, is on par with the national data.

- IV13 - school supports

Nova Scotia teachers report a very similar support provision model as New Brunswick and Newfoundland and Labrador – more supports than average from the school level (54% here, 51% nationally). Positive responses to the helpfulness inquiries were also more common with 'not provided' and 'not helpful' responses down from national scores.

- IV14 - division supports

There is proportionally slightly less division-level support in Nova Scotia (31% of reported supports) than is true across the nation (34%). They are not as well regarded as school supports since 44% of respondents answered that they were either 'not helpful' or 'not provided.' The highest rated response was that supports are 'helpful' (51%) with quite low numbers for 'very helpful' (4%).

- IV15 - ministry supports

Ministry supports are exactly on par with national data in terms of their proportions (15%). Ratings of their helpfulness are somewhat different though. Since 45% of respondents rated them as 'helpful' or 'very helpful' (the national figure is 44%) there are 5% fewer reports of 'very helpful' supports. And while negative responses ('not helpful' or 'not provided') total about the same proportions as the national numbers (54% in Nova Scotia, 55% nationally), respondents were 7% more likely to state supports are 'not helpful' and 6% less likely to state they are 'not provided.'

- Summary of supports data

In Chapter 3 it was indicated that Nova Scotia is the only province with **net** positive reactivity effects (when negative reactivity scores can, in effect, cancel out positive ones). It was also seen that it is also the second least reactive province overall (behind only Saskatchewan which currently has no province-wide assessment program). When looking at the supports that are provided for teachers in this province, especially in light of the fact that the jurisdictional provision of these supports follows a model common to three other provinces, these net positive and minimally reactive numbers are in some ways more confounding.

In provinces where most supports (as a proportion of the total available) come from the local level, there seem to be outliers, and then a core group of four (which includes Nova Scotia), which share many characteristics in the provision of supports and also how well-regarded these supports are. Alberta has 58% of provided supports at the school level and 89% of respondents consider them helpful or very helpful. British Columbia teachers have 50% of their provided supports from schools but only 35% consider them helpful at all (the 50% proportion might be misleading since 60% of BC teachers report supports are not provided by schools). These are outlying positive and negatives responses.

In the four provinces that remain in the school-based supports group, similarities in helpfulness are apparent. For all four provinces, the 'not provided'

and the 'not helpful' responses were the two lowest proportionally for Newfoundland and Labrador, New Brunswick, Nova Scotia, and Ontario. This indicates a fairly positive view of these supports. In all four provinces, the highest rated response was 'helpful' and the second highest for all was 'very helpful.' Again, school-provided supports appear to be well regarded by respondents. All these provinces also reported significantly lower ratings for both provision and helpfulness when looking at division- or ministry-level supports. Some divisions have, though, been active in providing both manpower and monetary supports for schools.

The provincial assessments gave us evidence enough to say, okay we need to put some additional resources in here, and we need to staff this. Both with resources in terms of books and literacy, you know, pieces as well as support with more literacy mentors and coaches in the schools to help the classroom teachers, as well.

- NS, Division staff, female

Because one of the big issues for us, and I'm sure this is everywhere, and that is schools tend to differ in their resource base even though we all work for the same board. . . So depending on the resources at the school, and often the size of the school determines that. The smaller schools are much more strapped for time and for resources for people to work with the children whereas the larger schools have a bit more flexibility. - NS, Division staff, female

With so many similarities in the supports factor, it is difficult to then account for the great differences in reactivity effects ratings. There is the most positively reactive province (New Brunswick) and the second least positively reactive (Ontario). The number one, two and three rated negatively reactive provinces (Newfoundland and Labrador, Ontario and New Brunswick) are next to the least negatively reactive province (Nova Scotia). The most reactive overall (Newfoundland and Labrador) neighbours the least reactive overall (Nova Scotia). Finally, the only net positive province is grouped with a net neutral (New Brunswick) and the second and third most net negative (Ontario and Newfoundland and Labrador).

It would be very difficult to reconcile these vast gulfs in reactivity by examining the supports variables alone. It is clear that assessment has been a focus area in Nova Scotia for some time, and perhaps this long-term focus has some influence on reactivity scores.

There has been a big push on assessment over the last 5, 6, 7 years, probably 10, even. . . We have certainly put a lot of time and energy into assessment. Now it always hasn't been around . . . the provincial assessment. It has been more around the classroom assessment, you know, that type of work that has gone on in the PD. And looking at classroom-based assessment has been a big deal and really having help teachers, you know, track the progress of their students on a regular basis. - NS, Division staff, female

The types of supports that are most frequently provided at the school level are: (a) administrative support (almost 80% coming from schools); (b) PLCs (62% from schools); (c) assessment teams (55% from schools); and (d) coaching/mentoring (51% from schools). Professional development was reported at a 40% provision rate from the school-level and printed or online guides were reported at only at 26% levels. This, of course leaves only one support most commonly provided by divisions (PD) and only one most commonly provided by the ministry (printed/online guides). These provinces are similar in more respects, but clearly the nation-wide reliance on schools to provide supports for teachers to use LSA data is evident.

Ontario

- IV12 - sharing of data (see **Figures 5.7 and 5.8**)

Ontario teachers report more data sharing than is true nationally, with the largest jump in the 'always' category (38% here to 24% nationally). The 'never' figure is quite low at 21% (the average is 31% nation-wide).

- IV13 - school supports (see **Figures 5.9 – 5.12, and 5.15**)

Most supports come from the school-level (as a proportion of all supports) and at the highest proportion in the country (59%). Ratings for 'helpfulness' are also highest in the country (56%) while 'not provided' responses were the lowest recorded in any province (2.1%).

- IV14 - division supports (see **Figures 5.9 – 5.11, 5.13 and 5.14**)

Ontario is a little lower than the national average for division-level supports (a 28% proportion compared to 34% nationally). They are also much less favourably received: the proportion of 'helpful' and 'not helpful' responses is equal at 37%. Still, 17% reported that these supports were not provided while 9% considered them 'very helpful.'

- IV15 - ministry supports (see **Figures 5.9 - 5.11, 5.14 and 5.15**)

A slightly lower proportion of teachers report ministry-level supports than is true nationally (13% to 17%). These supports are well-regarded (56% report 'helpful') by respondents.

- Summary of supports data

The most common supports in Ontario come from the school and these are also much more favourably rated than those from either the division (these are particularly poorly rated) or the ministry level. Interview subjects from this province had differing opinions about the effectiveness of supports in general. Note this comment on how seriously the graduation requirement Ontario Secondary School Literacy Test is treated by teachers:

The OSSLT has been around for at least 10 years in its sort of current form of a test . . . It has no bearing on anything in the school really and truly. I mean the principal is measured by this, teachers may be measured by it . . . there is nothing bad that happens if you're a bad teacher. And so the conversations, maybe they were in place five, seven years ago, but they're not anymore. The test is an eyeroll at best. - **ON, High school English consultant, female**

And then, from the same teacher and then a division-level employee, the insight that ideas about which supports could provide insights into test results differ:

Having rich discussions and connecting with literature about standardized tests or about teaching deeper and more enriching literacy activities, I think for them the literacy consultant was the best. I think for the majority of teachers . . . the progress that we saw was when we had people in the school . . . working with individual teachers . . . someone within the school, who knew the school's needs and tailored professional development to meet those needs, to close the gap. - **ON, High school English consultant, female**

'You are an under-performing school in the OSSLT,' or it is more typically in Math, 'we are going to bring you in centrally to Toronto.' All of those schools are going to come in with a team: the principal, some other key players, some other key teachers, and we are going to run you through high yield strategies, systematic approaches that we know work. That is the kind of investments that are made, and that is only at the provincial level. - **ON, Division staff, male**

The EQAO (Education Quality and Accountability Office) is arm's length from the ministry, but it seems that this distinction is not one that most educators make, specify, or fully understand. The provincial tests, testing policy, and the supports to use the data are all referred to under the umbrella term 'the ministry,' despite EQAO's distinct status. Certainly the assessment culture in Ontario is

strong and an expectation of the school system from the public at this point. Despite the fact that EQAO does not support the use of their assessment scores being used to rank schools, accountability is by necessity a public exercise and the wide-spread publishing of results in ranking tables is a sore point for teachers and education ministers alike.

“It is important to note that ranking schools solely on provincial assessments does not take into account all the range of factors that contribute to student success,” [Ontario Education Minister Liz] Sandals said in a statement. “The intent of the provincial assessment is not to rank schools, but rather to examine individual student, board and provincial data in order to improve instruction.” School boards use the Education Quality and Accountability Office (EQAO) tests to determine where there is the greatest need . . . adding the tests point individual schools and teachers in the direction of students who are facing challenges.⁸²

The Ontario model makes clear that school-wide and consistently low assessment scores are not a reason to curb funding – quite the opposite – they are a reason to provide more. Unlike the iconic American “No Child Left Behind” and “Race to the Top” programs, EQAO wants to help schools with external supports when scores are not good. This makes it more understandable, then, why Ontario teachers regard close-to-home, self-initiated, school-based support as a less intrusive and judgemental option than board- or ministry level 'interference.'

I would say that the least effective was professional learning communities [PLCs] and it is not because professional learning communities aren't effective because I really believe that they are. . . You are forced into professional learning communities and so sometimes they work, but when you are told you have to join and you don't like 3 or 4 of the other teachers, you're not feeling that sense of trust or willing to follow norms - you don't have an investment in the success of the group. And so I find professional learning communities, they can be the worst . . . When you are volun-told something? Absolutely.

- ON, High school English consultant, female

⁸² Artuso, A. (2014, February 2). Ranking schools against each other sparks debate. *The Toronto Sun*. Retrieved October 31, 2014 from <http://www.torontosun.com/2014/02/01/ranking-schools-against-each-other-sparks-debate>.

School-level supports, while well regarded, can certainly be said to be the most variable, the most administrator-driven, and the least consistent by nature of the small scale on which they are provided. There are differences in opinion about whether a tight or loose framework is the most productive.

When I do the [school-based] PD session on the PD day between first and second semester, it is not optional. I basically say, like . . . we are going to do the grade 10 literacy prep in period 1, grade 10 classes . . . And I say, 'You know what? This is what we are doing and I need you to be aware of that, that one day a week for the next, you know, the first six or seven of the semester you are going to take the bulk of the class to work on this. And we're going to do a PD session just to be sure everyone knows what the lessons look like. . . . This is way it is going to be. This is the lesson plan. This is the lesson plan that has worked for the last dozen years - do not deviate.'

- ON, High school principal, male

The best answers come in variety, so we [at the division level] are loathe to funnel our folks into specific strategies. We want to give them the freedom to create success in their own particular context, their own way. When they struggle, we provide them with possible models. . . we think that by allowing them to work with other schools or bring all the schools, typically, more often, to share how they do things. - ON, Division staff, male

Prince Edward Island

- IV12 - sharing of data

PEI respondents report some of the lowest levels of results sharing in Canada. A 41% proportion report never sharing results and 38% report only sometimes sharing these data (national the numbers are 31% and 45% respectively).

- IV13 - school supports

PEI shows the most equitable distribution of support provision across the three jurisdictional levels. This may be a consequence of the relatively small area of the island itself. Most supports do come from schools, though (40% of the total) and are well regarded (72% of respondents had positive ratings). The 'not provided' response is higher than national numbers by 7% (24% to 17%).

- IV14 - division supports

Divisions provided 37% of the total supports and this is slightly more than the national proportion (34%). They are better regarded than national data as indicated with 70% positive ratings compared to 58%.

- IV15 - ministry supports

The proportion of ministry supports is 8% higher than the national average, but these supports are not particularly popular. An equal level of 25% report not helpful supports from the ministry and 25% report they are not provided from the ministry-level. This leaves the other 50% of respondents in the positive rating groups 'helpful' (31%) and 'very helpful' (19%).

- Summary of supports data

Chapter 3 looked at this study's reactivity results without trying attribute these effects to any root causes. Moving forward through independent variables, it is becoming more clear which of the variables examined in the teacher survey are most closely tied to total reactivity, and which to either positive or negative reactivity effects. Prince Edward Island came out of that preliminary analysis as the third least reactive province overall (behind Saskatchewan and Nova Scotia), and the second most net neutral (cancelling out positive effects with negative ones to leave a 'net score' close to 0) behind only Saskatchewan.

The supports data show that what is provided for teachers comes from all jurisdictions fairly equitably and is well regarded from school and division levels. The least commonly provided and least likely to be appreciated are ministry-level supports, although there are some indications things are changing for the better.

Our PD days are all school-focused and school-based, instead of like somebody coming in from the [education] department to teach all the grade 3 teachers about math. That is what we're getting next year but that is kind of, that's all a result of all the PISA foolishness.

- PEI, Elementary homeroom teacher, female

But yet when the math scores are not where they need to be there are fewer numeracy coaches and we haven't had a numeracy coach come through the doors in the last three years. . . I think some of it has to be some of the support at that higher level to come and help support and identify. - PEI, High school Math teacher, male

I have to say, though, at first, it was terrible. They would come in and do the 'dog and pony' slide show and hand you the results and they'd leave. Now we work with the results, it is a little different now. People, I guess too, people are more confident, they know what questions to ask, they know what they are looking for.

- PEI, K-9 school principal, female

We have curriculum coaches now, umm, literacy coaches and numeracy coaches who have been put in place, not enough I must

say, to help with the work. . . It's great they are kind of intermediaries, they take some time and look at the results and help your school development team or your school effectiveness team work. And they get into individual classrooms as well, so the board does take this seriously and the focus is on improved instruction, for sure. - **PEI, K-9 school principal, female**

The Prince Edward Island assessment program may be regarded as a little long-in-the-tooth by teachers, and in need of some renewed emphasis and attention. There were also cases where the ministry responses to assessment results were not thought to be very effective.

When we began the literacy assessments six, seven, eight years ago. . . as the results came back, you know, there was kind of that awareness that, 'oh goodness look! . . . Let's give them all a workshop on this' or whatever. . . That has kind of dwindled because funding has. Everything just gets less and less and less. Five years ago I would have had five workshops in a year. This year I think I had two. - **PEI, Elementary homeroom teacher, female**

Sometimes, yeah, I think, 'Is the tail wagging the dog or the dog wagging the tail?' sometimes. And I think sometimes in language arts, that is it. . . I guess it is the approach, what happens after [the LSA], and what schools do individually, and then what the board does at their level, and then provincially. . . I'm sure I had a couple of moments where I thought, 'Oh, my god. And this is the response? You know, get serious!' - **PEI, K-9 school principal, female**

Teachers in many jurisdictions report a similar situation where assessment programs were a priority when they were introduced but with a never-ending onslaught of new initiatives (ask any teacher about this) they seem to have been supplanted as top priorities and appear to have been relegated to a smaller role in the day-to-day workings of schools. That cannot be a good thing for policies goals that strive to make the data relevant to teachers and parents.

“. . . our national and international assessments tell us that we have work to do in our education system in PEI and our provincial assessments let us know where this work should happen. . . And what teachers say to us, particularly from being on the marking

boards, is that they have changed their practice based on what they see from the assessments.”⁸³

Despite the fact that 59% of PEI respondents reported being participant in marking and/or writing items (i.e. marking boards), the changes in practices were not widely reported, and were more often negative reactivity effects than positive ones. There are definitely policy-level benefits to having data from schools (while not necessarily census-style data collection), but school-level benefits depend on the data being considered a priority and being used by teachers who have the training and motivation to use them appropriately.

I mean, just the test-giving itself requires a little bit of training. . . The work that comes after, umm, they have a marking board so that teachers have training on the marking board to see how the results are done. And then, umm, I would say for the most part there's, especially in math, there's all sorts of project work with teachers in groups. . . to help teachers deal with the curriculum and the shortcomings. - PEI, K-9 school principal, female

Québec

- IV12 - sharing of data

Data are shared by teachers in Québec at very similar rates to the national ratings for this metric. Most report they are shared 'sometimes' (48%) while 29% report they are always shared.

- IV13 - school supports

Québec has proportionally fewer school-level supports than any other province (39% compared to the national average proportion of 51%). That said, the helpfulness rating for school supports are very much in line with national averages with only two significant differences: the proportion of respondents rating school supports as 'not helpful' is lower than the national average (3% to 7%) and the proportion of teachers indicating supports were not provided is higher than the national average (24% to 17%).

- IV14 - division supports

Québec is the only province that reports more division-level supports than school-level ones (49% to 39% as proportions of the total). These supports are also highly rated for their helpfulness – 68% find them helpful or very helpful.

⁸³ PEI Department of Education and Early Childhood Development. (n.d.) Provincial Assessment Program. Retrieved Sep. 5, 2014 from: <http://www.gov.pe.ca/eecd/studentassessment>

- IV15 - ministry supports

Ministry-level supports are somewhat less common than in the national picture (12% as a proportion of the total compared to 15% nationally). Their ratings for helpfulness are quite low since 37% of respondents indicated they are 'not helpful' and 44% stated they are 'not provided.' Both of these scores are above the national averages (20% and 36% respectively).

- Summary of supports data

As the province with the second highest average score for total reactivity, it is clear that teachers in Québec do use LSA data to inform their instructional practices. The balance is toward negative reactivity, like most other provinces. In Québec's case, the supports that are provided come from a different source than what is more commonly the case. Québec is the only province to get most supports from the division level. This might be seen as an effective way to ensure that more teachers get a common message and thus are more likely to move in the same direction than if schools are left to direct (and interpret) what ministry goals and policies mean for them individually.

Examining the reactivity scores, though, it is not clear that division-level supports are any more effective than school-level ones at directing reactivity effects to the positive side. It might be that the divisions are also impotent to alter the practices and the secrecy built into ministry assessment policy.

So we have our regular math textbook and if you follow the progression of learning they really, really, specifically tell you what you need to look at. But even at the school board level, they don't even know what the exam is going to be like year after year until the last minute. - **QC, Middle years homeroom teacher, female**

Last year I was teaching grade 5 and of course we also give a situational problem at the end but we don't have the marking centre [mandatory marking meetings]. So we can give it maybe a little bit before, we have that freedom. But it is still a lot to correct, it is crazy at the end. So I can't even help my colleague that will be in grade 6. This year I will be in grade 6 and I don't expect anybody to help me out. - **QC, Middle years homeroom teacher, female**

The sense of frustration with ministry policy-making opacity was apparent in discussions with teachers from Québec, and no one I spoke to seemed to be on solid footing talking about assessment policy on the large scale – they know what was true for their school, and that was about as far as they could speak in an informed manner.

We never see [results data] - it is very rare. We see them maybe at the beginning of the next year . . . presented by the principal and sometimes we have the consultants come here to show us. . And we try to work out where we went wrong and what we can do to improve for the next year.

- QC, Middle years homeroom teacher, female

PLCs [professional learning communities] in our board have not trickled down beyond the level of board administration. So the administrators are in PLCs, or what they call PLCs - I don't personally believe they are PLCs. They are trying to generate PLC-type behaviour. . . [using] social networking for teachers, so they'll set up for like language arts in our board, they'll set up for various things, and they want teachers to collaborate across those networks.

- QC, High school English teacher, male

It appears that the education ministry in Québec has not been particularly effective in either creating a culture amongst educators where provincial assessments are considered useful or sharing a positive message with educators about assessments through provision of supports. The very low ratings given to provincial supports here speak to that failure.

Saskatchewan

- IV12 - sharing of data

Very much in line with national responses, Saskatchewan shows 47% of respondents getting data shared 'sometimes,' 27% 'never,' and 26% 'always.'

- IV13 - school supports

The proportion of supports coming from schools is 5% less than what is true nationally, but these supports are rated to be just about as effective and helpful. Just 10% of respondents rated school supports as 'not helpful,' and 17% rated them as 'not provided.'

- IV14 - division supports

There are proportionally 6% more division-level supports in Saskatchewan. Their ratings for helpfulness are not vastly different than national figures in this metric, but they are considered slightly more helpful (a 2% higher rating for 'very helpful,' and a 4% higher rating for 'helpful.').

- IV15 - ministry supports

Ministry supports were rated, as a proportion of the total number of supports, just about on par with the national average (rounded to 15% for both). Yet 57% of respondents indicate that ministry supports are not provided – the second highest

rating for this metric. At 31%, the 'helpful' rating is not high, and the 'very helpful' rating does not even register.

- Summary of supports data

At the risk of being repetitive, Saskatchewan alone has an assessment program that does not currently affect most teachers or most schools. New assessments are being piloted and will certainly be introduced at other grade levels in the future, but the respondents in this survey were relating their experiences with a now discontinued battery of tests called the Assessment for Learning (AfL) program. That makes the lessons gathered here irrelevant for *reform* of existing policy, yet very important to *inform* the policies that are being and will soon be devised.

The ministry stands out in these data as the jurisdiction that did not effectively deliver their message to teachers. Too many respondents stated that ministry supports were not available to help them work with assessment results. When divisions and schools are left to fill in this void, the policy can suffer from a lack of 'fidelity:' more varied interpretation as a result of more interpreters (Spillane, Diamond, Burch, Hallett, Jita & Zoltners, 2002; Elmore, 1980; Anderson, Leithwood & Strauss, 2010). Despite the fact that Saskatchewan teachers had a fairly respectable net reactivity average score, both the amount of reactivity and the total reactivity average were very low. The messages coming from the ministry and school divisions about AfLs were not clear to staff.

There wasn't PD [professional development] on it. There wasn't anything. I don't think so, there wasn't any PD or anything. We were just given it, it was given to us, go ahead and use it and do it, show it to your students. . . . And a little bit of collaboration with the other grade 7 teacher to say 'let's do it on this day.'

– SK, Middle years homeroom teacher, female

As a principal, I felt that their [the school division's] idea of 'support' was I had to get the scores up. And that was, so it was maybe not so authentic as I would have liked it. Because I wasn't really hung up at improving the scores, at a school level, I was kind of more hung up on being more responsive to needs of all the kids in the school.

- SK, Elementary school principal, male (a)

[Discussions of data] were pretty brief, I think and kind of gone over and . . . Briefly, and in saying that we do sometimes have discussions a little bit to say, 'oh, this is interesting' or not, but there is not really a lot of follow up on what to do next.

– SK, Middle years homeroom teacher, female

Any new assessments will need to address these shortcomings. Reactivity should be improved (if only to meet policy intentions), and positive reactivity should be emphasized. Saskatchewan does not have the high stakes diploma exams, provincial exams or minimum competency exams that are common in many provinces (excluding the few classes that write departmental exams – see Chapter 1). This factor is important in that there is less pressure on teachers (as will be shown in Chapter 6) to help students achieve at all costs. LSAs should, ideally, drive positive reactivity and the specific policy features that will be coming in Saskatchewan will determine if this takes root. Some suggestions from the field include:

Oh yeah, people got the money to provide resources in that area and then spent it on something else. Technology, you know, elaborate PD. You don't need to send anybody to ASCD (Association for Supervision and Curriculum Development) in San Antonio . . . We are sending people all over North America to figure out how to teach. If you come to my acreage I can help them out. . . Don't get me wrong, I think there is value in that, but I think we really have to start seriously looking at . . . there is too much of a discrepancy in PD . . . we have to give everybody a chance at some level.

- SK, Elementary school principal, male (a)

I think [teachers like PLCs] because they get to talk. They are smaller groups and they are more focused to their teaching assignment rather than it just being a general session quite a number of people. It is focused by grade and subject area, so it is a little more specific. They can get more in depth as to how they are using assessment in their classrooms. - **SK, Elementary school principal, male (b)**

5.5 Correlation analysis – supports

The independent variables in this section are examined using Spearman's rank order correlation tests in order to determine relationships that exist between them. Significant relationships will be indicated with an asterisk for significance at the $p < 0.05$ confidence level and two asterisks for significance at the $p < 0.01$ confidence level.

Examining the independent variables first (see **Table 5.2**), it is noticeable that most of the variables in this chapter are positively and significantly correlated with the others with the notable exception of ministry-provided supports. The sharing of data, for example, is positively and significantly correlated with school- and division-level supports as well as their considered helpfulness. The highest

levels of correlation are between school- and division-level supports being provided and the respondents' opinions on the helpfulness of the supports. These are correlated at highly significant levels (0.413 and 0.479, respectively). The provision of ministry-level support is also significantly and positively correlated to helpfulness, but this is the only significant correlation for this variable. The levels of correlation, while high, do effectively exclude collinearity as a concern in the regressions that follow.

Table 5.2: Spearman's rank order correlation test done with supports variables

Correlation matrix - supports variables

1. Sharing of data	1.000				
2. School supports	0.283**	1.000			
3. Division supports	0.268**	0.175**	1.000		
4. Ministry supports	0.014	-0.016	0.092	1.000	
5. Helpfulness of supports	0.379**	0.413**	0.479**	0.184**	1.000

* p<0.05; ** p<0.01

5.6 OLS regressions – supports variables

5.6.1 Regression analysis

Reactivity is the dependent variable in this study, yet it has (as seen in Chapter 3) both positive and negative effects. There are two other elements in reactivity derived from the positive and negative results. Total reactivity adds the absolute values of positive and negative effects together to measure how reactive in total teachers are to the LSA results data. Net reactivity subtracts negative effects from positive to determine the overall balance between positive and negative effects. In the regressions that follow, since net reactivity is based upon mathematically cancelled out values, it will not be examined. The other reactivity options will be discussed in this order: positive reactivity; negative reactivity; and total reactivity.

The tables show the coefficient in the first row and the *t* statistic in the second row (significant relationships are indicated here using an asterisk). Provincial dummies were added to examine variations at this level, and PEI is the control province (it does not have dummy added) for total reactivity, BC for negative reactivity, and MB for positive reactivity.

Table 5.3: Positive reactivity is seen here in light of supports variables.

Positive reactivity

Sharing of data	0.180 (3.59)**	0.174 (3.38)**
School supports	0.008 (0.24)	-0.010 (0.29)
Division supports	0.138 (3.35)**	0.141 (3.40)**
Ministry supports	0.069 (1.19)	0.058 (0.98)
Helpfulness of supports	0.075 (1.70)	0.070 (1.55)
(provincial dummies) AB		0.494 (1.85)
BC		-0.145 (0.50)
NB		0.366 (1.41)
NL		0.263 (0.90)
NS		0.169 (0.64)
ON		0.295 (1.09)
PEI		-0.218 (0.84)
QC		0.247 (0.85)
SK		-0.555 (1.90)
Constant	2.280 (16.38)**	2.215 (9.25)**
R²	0.15	0.22
N	338	338

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

There are two variables that have strong correlations to **positive reactivity** effects (see **Table 5.3**). These are the **sharing of data** (which includes both the respondent sharing with other teachers and also other teachers sharing with the respondent) and **division-level supports**. These variables remain significant prior to and following the introduction of provincial dummy variables. The sharing of data is perhaps the most basic measure of a data-informed school culture, and as such should be an indicator that positive reactivity practices are common. A school staff that has the comfort level to share data as well as the skill set to make sense of them is in a good position to improve their instructional practices based on that footing.

The second significant variable is something of a surprise. Divisional supports trailed school supports in almost all provinces for their quantity (how often they are made available) and for their reported helpfulness. Yet it is division-level supports that have the close tie to positive reactivity. School and ministry supports were both found to have insignificant effects. This may be since divisional supports are much more common than ministry level supports and have a firmer grounding in both data-analysis skills and instructional methods than would be true of most school-level supports. It should be said that expertise does not always flow from central offices, but it is likely the case that a teacher tasked with helping the division reach a literacy goal, for example, has access to more resources and has demonstrated a more varied set of instructional skills than most English teachers at the school level. For these reasons, division level supports along with school-level data sharing are responsible for 22% in the variance related to supports and positive reactivity. None of the provincial dummies show significant deviations from the control group.

Examining **negative reactivity** in light of supports, there are no significant correlations after the provincial dummies are added to the regression table (see **Table 5.4**). Only one significant relationship exists prior to this addition and that is where data are shared between teachers, there is also more negative reactivity. This negative correlation is the only one for independent variables that appears in this table, and it follows closely upon what we have just seen with positive reactivity: when data are shared, the tendency for teachers is to move towards reactivity effects, and to shun neutrally reactive practices (inaction).

This certainly makes clear the fact that the sharing of data is a key driver of instructional change. The amount of variance that is explained by this one variable (before provincial dummies are added) is not large at 2%.

Looking at the provincial dummies, there is some variance in responses between jurisdictions from the control group. Both Saskatchewan and Nova Scotia show highly significant positive correlations and Manitoba shows a less significant positive correlation.

Table 5.4: Independent variables related to supports are examined for negative reactivity effects. **Note that since negative reactivity is counted in negative integers, a negative coefficient means more negative effects.**

Negative reactivity

Sharing of data	-0.127 (2.22)*	-0.074 (1.33)
School supports	0.032 (0.80)	0.066 (1.77)
Division supports	-0.029 (0.63)	-0.056 (1.26)
Ministry supports	-0.080 (1.22)	-0.116 (1.83)
Helpfulness of supports	0.042 (0.83)	0.022 (0.47)
(provincial dummies) AB		-0.146 (0.52)
MB		0.672 (2.22)*
NB		0.063 (0.23)
NL		-0.169 (0.55)
NS		1.124 (4.02)**
ON		0.536 (1.80)
PEI		-0.399 (1.44)
QC		-0.088 (0.29)
SK		0.963 (3.13)**
Constant	-3.151 (20.00)**	-3.379 (13.40)**
R²	0.02	0.19
N	339	339

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

Positive correlations indicate that less negative reactivity is apparent in these three jurisdictions than that which is seen in the control province (BC). As the R^2 value becomes quite large with the addition of provincial dummy variables (it rises to 19%), insights may come from a closer examination of the supports-related independent variables in terms of the provincial variation. With so much of the variance explained by provincial differences, this appears to be the most telling aspect for negative reactivity.

Turning to **total reactivity** results (see **Table 5.5**), they align quite well with the positive reactivity results despite the lack of significant correlations within the negative reactivity data set. These variables did not show significant relationships to negative reactivity, but the results in **Table 5.4** show that they do increase negative reactivity effects, if only at non-significant levels. In conjunction with the strongly significant positive reactivity result, both **sharing of data** and **divisional supports** variables create the necessary conditions for reactivity.

There are two provincial dummies with highly significant negative correlations which indicate a high level of divergence between the results in both Saskatchewan and Nova Scotia with the control group (in this case, PEI). This result, in contrast with the independent variables, seems to be driven by negative reactivity effects. Both Nova Scotia and Saskatchewan reported very low levels of negative reactivity (the lowest levels nationally are for these two jurisdictions, as seen in Chapter 3). It can be concluded that the relatively low levels of negative reactivity reported in these two jurisdictions are the reason the results diverge from the control group (since the positive reactivity regression showed no such provincial variations at significant levels).

In terms of the correlation matrix examined above (**Table 5.2**), these regressions indicate that while supports, sharing and helpfulness were strongly correlated with one another, this did not translate into these variables being significant in terms of reactivity effects. In hindsight, it may seem telling that the highest level of correlation was between divisional supports and helpfulness. It is clear that the provincial data examination above clearly indicates that divisional supports were not as well received as school-based supports, and these numbers might be inflated by the number of respondents who indicated that divisional supports were not available. Providing these supports, which apparently are very effective at promoting reactivity, should be more of a priority.

5.6.2 Residual analysis

The residuals from these regressions were examined using four different econometric graphing techniques and the results from these analyses were fairly uniform across all three regressions seen above. The results for the **positive reactivity** residual examinations are found in **Figure 5.6**.

Table 5.5: Total reactivity effects correlated against supports variables.

Total reactivity

Sharing of data	0.308 (3.56)**	0.259 (2.99)**
School supports	-0.015 (0.24)	-0.065 (1.12)
Division supports	0.177 (2.51)*	0.201 (2.89)**
Ministry supports	0.137 (1.38)	0.161 (1.63)
Helpfulness of supports	0.035 (0.45)	0.048 (0.63)
(provincial dummies) AB		0.435 (1.16)
BC		-0.263 (0.61)
MB		-0.768 (1.75)
NB		0.170 (0.46)
NL		0.247 (0.59)
NS		-1.192 (3.06)**
ON		-0.398 (0.95)
QC		0.164 (0.38)
SK		-1.693 (3.93)**
Constant	5.412 (22.67)**	5.744 (17.84)**
R²	0.10	0.21
N	331	331

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

Results for the other two regressions are found in the chapter-ending section. The 'observed v. predicted values' chart shows in this case a positive linear trend. A fairly clear linear trend is a good indication that the regression model has some validity. The bands seen in the graph indicate different self-reported levels of reactivity from survey respondents (from 0 to 10 by multiples of 0.5).

The ' \hat{e} v. \hat{y} ' chart has these same bands, but does not indicate clustering or the presence of serious outliers. Apparent clusters or outliers in this chart could portray the regression coefficients as less accurate in light of their ability to distort regression coefficients.

The residual histogram has a quite normal distribution. Although not perfectly in line with the normal distribution curve, the residuals are more likely to meet the assumption of normality and independently distributed residuals required for hypothesis testing using OLS methods when they are this close to that standard.

Finally, the QQ plot indicates a fairly normal distribution at the tails. This quantile analysis checks the tails of the distributions, and in this particular case, very little deviation from the normal distribution is apparent.

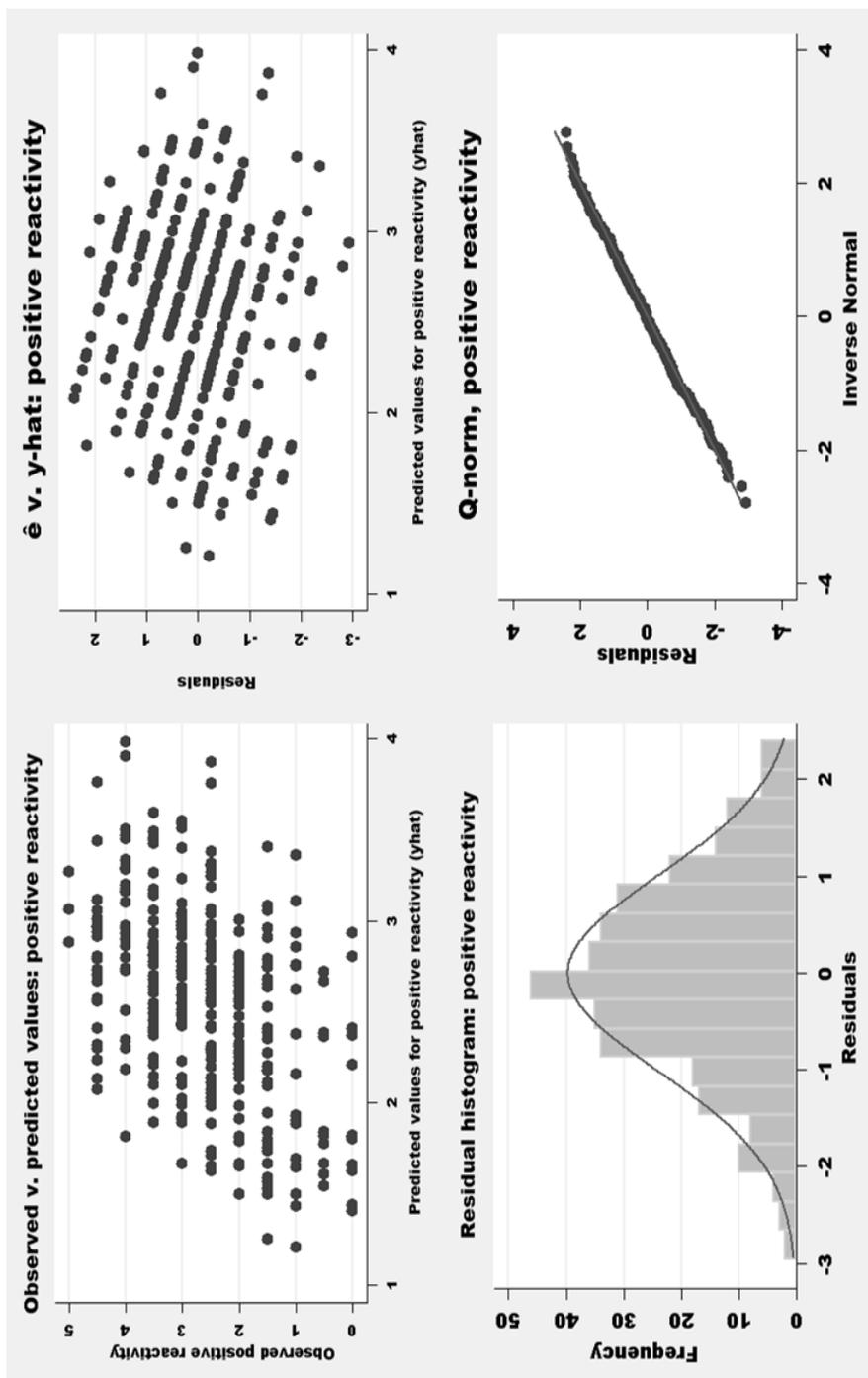
These analyses are very typical of the residual analyses found in the chapter-ending sections of all the central chapters. There is some weak clustering visible in **Figures 5.16** and **5.17** at the end of this chapter. And while the data do show some deviation from normal distributions as a result of the metrics used, they do not contradict the quality of the data set or provide different interpretations of the data. The residual analyses thus add some rigour to the statistical model.

5.7 Conclusions

The hypothesis stated as the basis of this chapter should not be rejected outright since with some caveats attached, it still does have some predictive value.⁸⁴ It seems not to be relevant (in terms of reactivity effects) whether or not respondents found the supports provided helpful, and it does not seem to be the case that more supports lead to more reactivity. This analysis indicates that supports are most effectively provided from a jurisdictional level high enough to access more expertise and more resources than school-level supports can.

⁸⁴ **H 5-1:** Teachers who feel supported in their professional practices will be more inclined to adopt data-informed instructional techniques. Thus, positive opinions about provided supports and the recognition that supports are available will have a positive impact on **total reactivity** scores and more specifically increase **positive reactivity**.

Figure 5.6: Residual analysis for supports data and positive reactivity regressions



More school supports (there are already more of these than divisional supports) would not, according to these results, increase reactivity or alter the reactivity practices of teachers.

That said, narrowing the focus of the provided supports to the division level and negating the significance of reported helpfulness, the second part of the original hypothesis holds true: division supports are significantly correlated with total reactivity and even more strongly correlated with positive reactivity.

The sharing of data between teachers relates to the opening statement of the hypothesis, that teachers who feel supported are more likely to adopt data-informed practices. This is just about equally significant in these results as divisional supports. The sharing of data is a school-level platform that allows educators to pool their strengths and move their instructional practices forward. It is worth acknowledging that while not exactly a school level *support*, a culture of sharing and working with data is most commonly the effect of a school leader who sees the value in such practices (as noted in Schildkamp & Kuiper, 2010; Weinbaum, 2009). Sharing data with colleagues is not without benefits, and also not without personal risks. These risks have to be well navigated to make the positive benefit clearly outweigh the potential drawbacks for staff members. Note the comment by a New Brunswick administrator about a school culture that both permits and expects sharing:

We have a 'community of trust', and if we don't have that, we have problems. So I think all of those formative tests and assessments . . . allow us that opportunity for accountability for teachers, accountability for curriculum, but also accountability for sharing and professional development. - **NB, High school principal, male**

The importance of the school-based administrator in creating such a culture is vital to the successful implementation of ministry or divisional policies.

One final point to note is that there are no significant links between supports variables and negative reactivity. This is an encouraging result. If negative reactivity effects were correlated with any kind of professional development or support at any level, this would indicate that the in-servicing being provided to staff was focused on those strategies that have the least educational leverage.

5.8 Charts and tables

Figure 5.7: Respondents rated both themselves and teachers from 'feeder' classes (their current class' teacher from the last school year) on whether these data were made available all the time, some of the time, or never. There are large variations between provinces and the sharing of data between teachers/grades appears to be inconsistent at best. The national data also indicate a higher proportion of 'never' than 'always' responses.

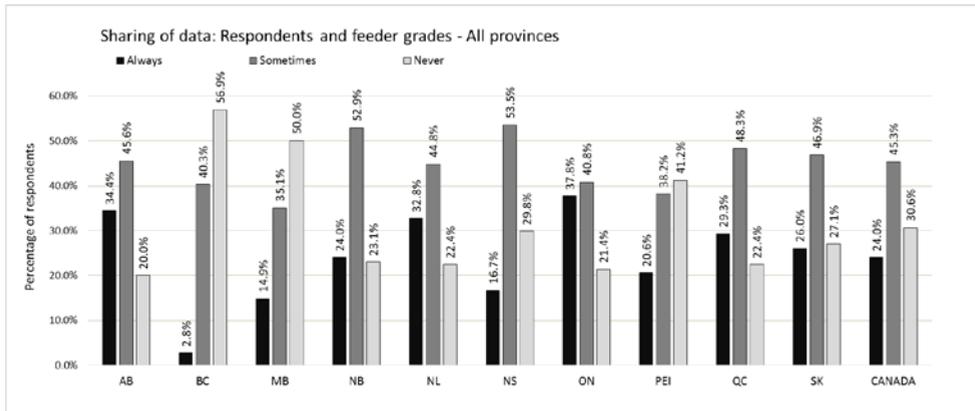


Figure 5.8: Using the national numbers from figure 5.7, this chart compares them with responses from only those teachers who do not give LSAs but get students coming from classes/teachers who do. This might be a student, for example who writes a grade 6 Math LSA in one year, and the next year is in a respondent's grade 7 Math class. The number of respondents in this category indicated significantly lower rates or results sharing than teachers who themselves give LSAs.

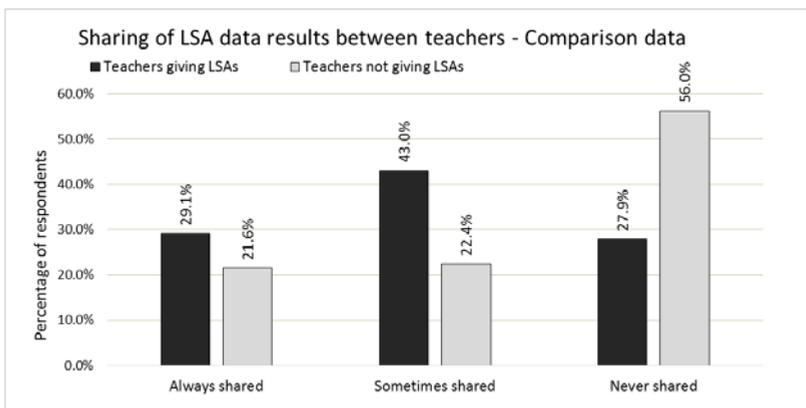


Figure 5.9: Six types of supports were mentioned most prominently in the literature and data for each are shown here differentiated by which of the three jurisdictional levels provided respondents with that specific support. Schools lead in the provision of most support categories except for professional development (which is more costly to provide for a small group than a large one) and printed or online guides (which would more generally come from the ministry that creates or oversees the assessments). Divisional support is apparent across the spectrum, while at generally about half as common as in-school provision. The relative invisibility of ministry supports is quite striking. Aside from printed or online guides, they have a presence of 10% or less in 4 of 5 categories.

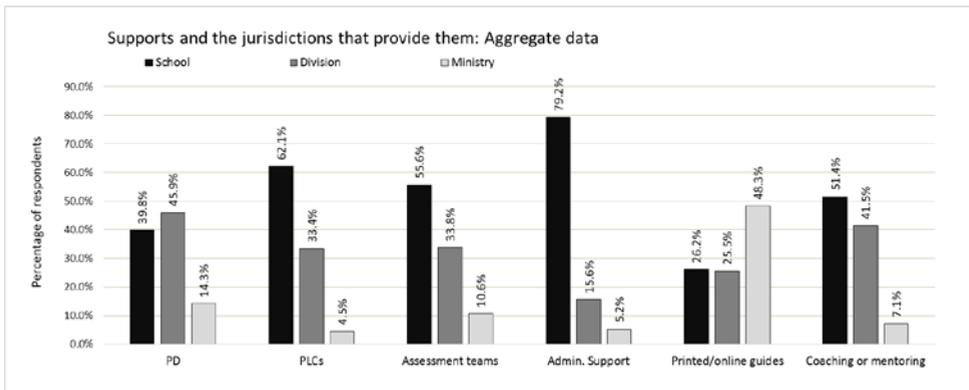


Figure 5.10: This chart shows which types of supports were provided to respondents as prompts and which were most commonly reported.

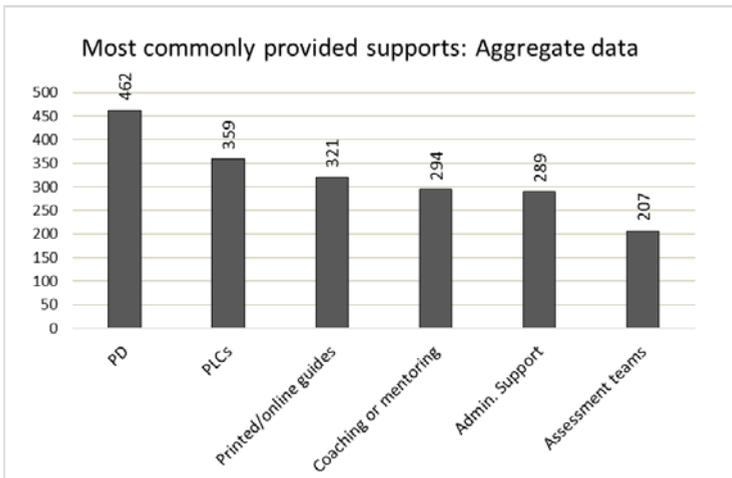


Figure 5.11: This chart compares which jurisdictions lead in providing supports. In all provinces except Québec, the school takes the lead in providing supports for teachers to use LSA data. In Québec, it is divisions that lead (48.8%) by a margin of 10% of respondents over schools. The only provinces where more equity is shown between schools and divisions are Québec, Prince Edward Island, Saskatchewan and Manitoba. The ministry is the least visible jurisdiction in providing supports, indicated by less than 10% of respondents in two provinces, less than 20% in six more.

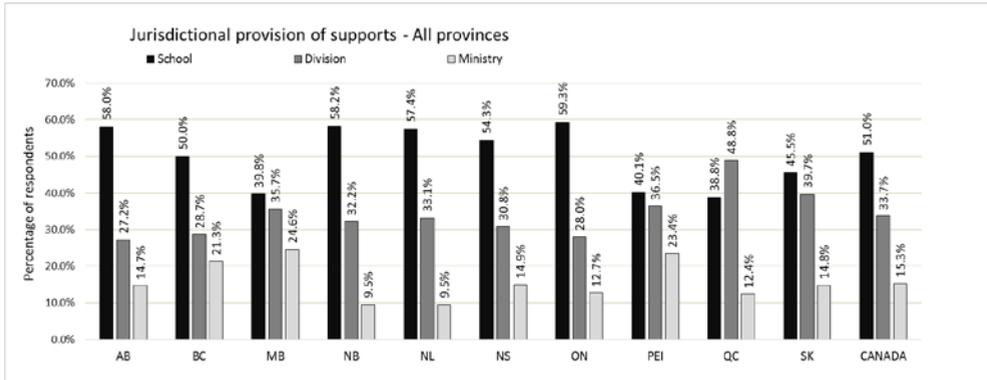


Figure 5.12: Most respondents indicate that school-based supports are not only provided, but also helpful in the use of LSA data to guide and improve instruction. British Columbia is an outlier in this regard where only 34.3% of respondents found the supports 'helpful' or 'very helpful'. In all other provinces the percentage for the two positive categories (helpful or very helpful) was 70% or higher. The figures for 'not helpful' and 'not provided' supports are still significant, though, running 6.8% nationally for 'not helpful' and 17.2% for 'not provided'.

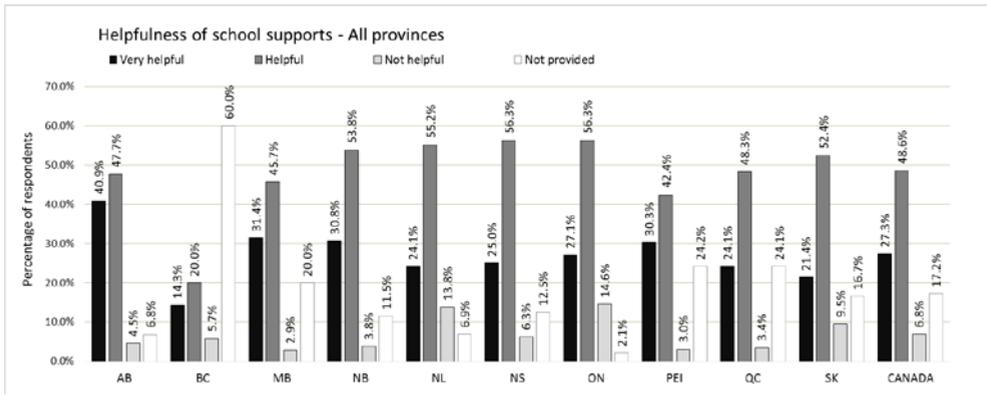


Figure 5.13: Divisional supports, while less commonly provided as those from schools, were not nearly so much appreciated. 'Very helpful' numbers dropped from 27.3% to 9.9%. 'Helpful' responses were slightly off from 48.6% to 47.5%, and the majority of the difference between positive and negative responses was made up in the 'not helpful' and 'not provided' categories: from 6.8% to 17.9% and from 17.2% to 24.8%, respectively.

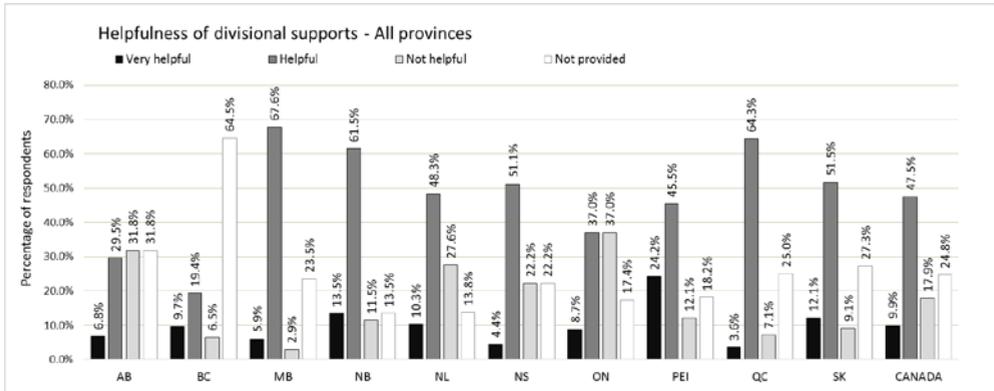


Figure 5.14: It appears that as the jurisdiction gets more remote from teachers (becoming less local), fewer supports are provided and they are also considered less helpful.. 'Very helpful' and 'helpful' numbers appear respectable (7.1% and 37.1% respectively) except in light of the proportion of negative responses. 19.9% of teachers found ministry supports to be 'not helpful.' And fully 36% of respondents did not indicate they had any ministerial supports provided. Together, these negative responses make up 55.9% of the responses in the sample. This compares to 42.7% with negative responses regarding divisional supports and only 24% with negative responses about school supports.

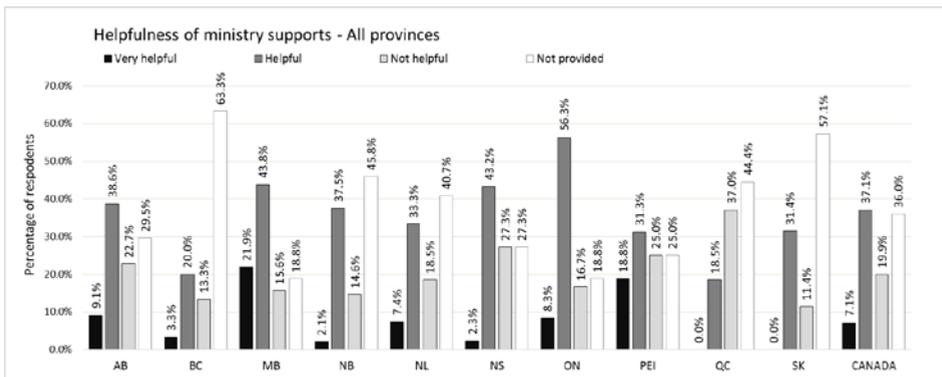


Figure 5.15: This chart shows the relative value placed on provided supports by teacher respondents across Canada. The first two bars, helpful and very helpful, decrease in relative size as you move from the local to the central (school, division, and ministry). The last two bars, which are the negative responses, not helpful or not provided, increase in relative size as one moves in that same direction. While it is more time and cost efficient to provide some types of supports to large numbers of teachers, the supports that are most common and most commonly considered helpful are those found close to home, and provided in-house by the school. The strong link between division-level support and reactivity effects means that these types of support (be it coaching, professional development, or PLCs) provide more 'bang for the buck' in delivering positive instructional change based on LSA data.

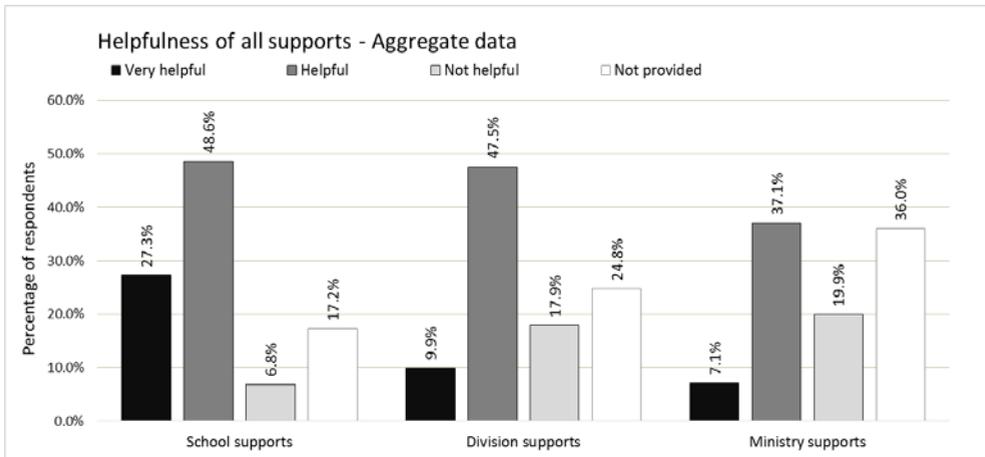


Figure 5.16: Residual analysis for supports data and negative reactivity regressions

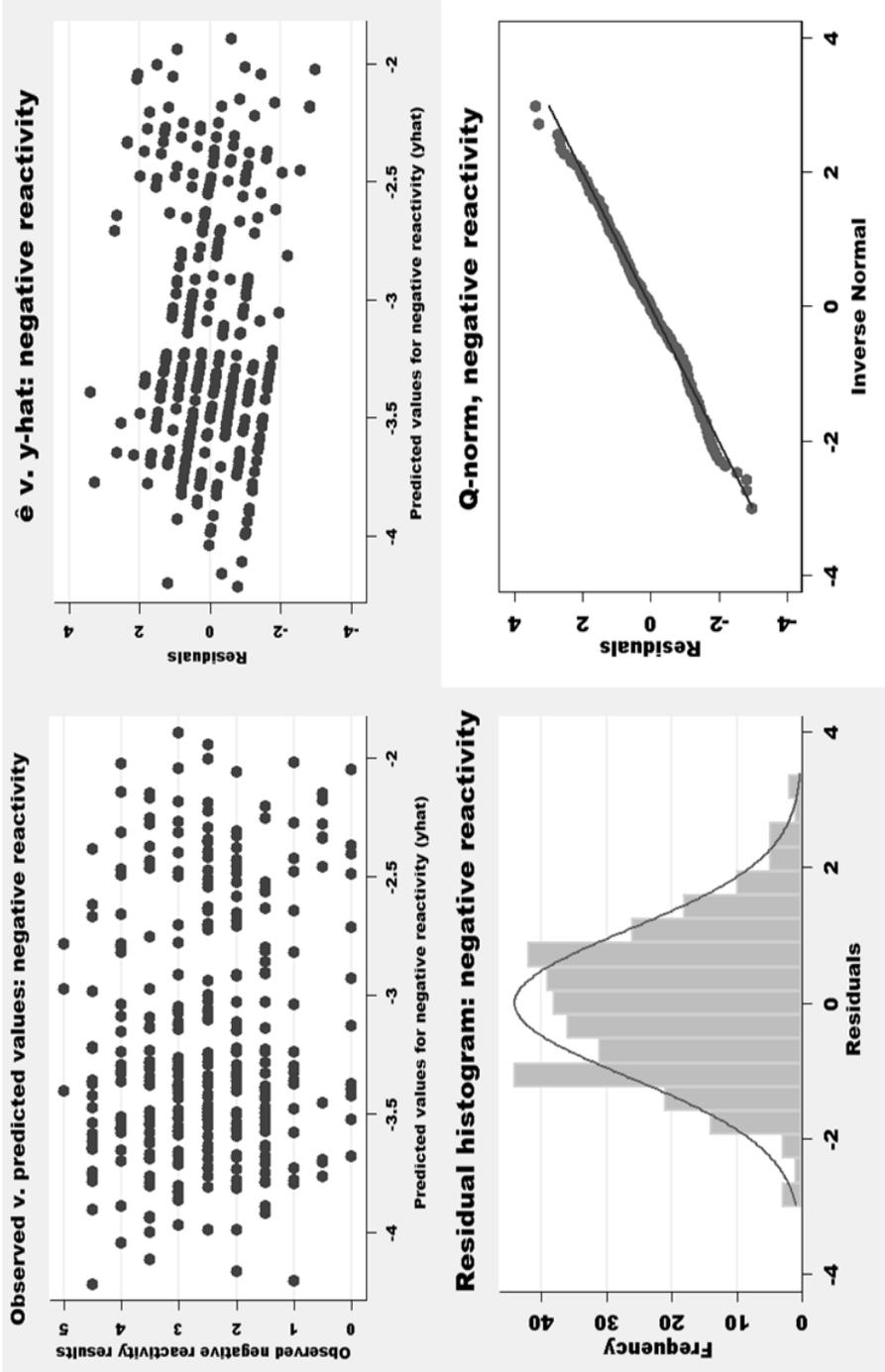
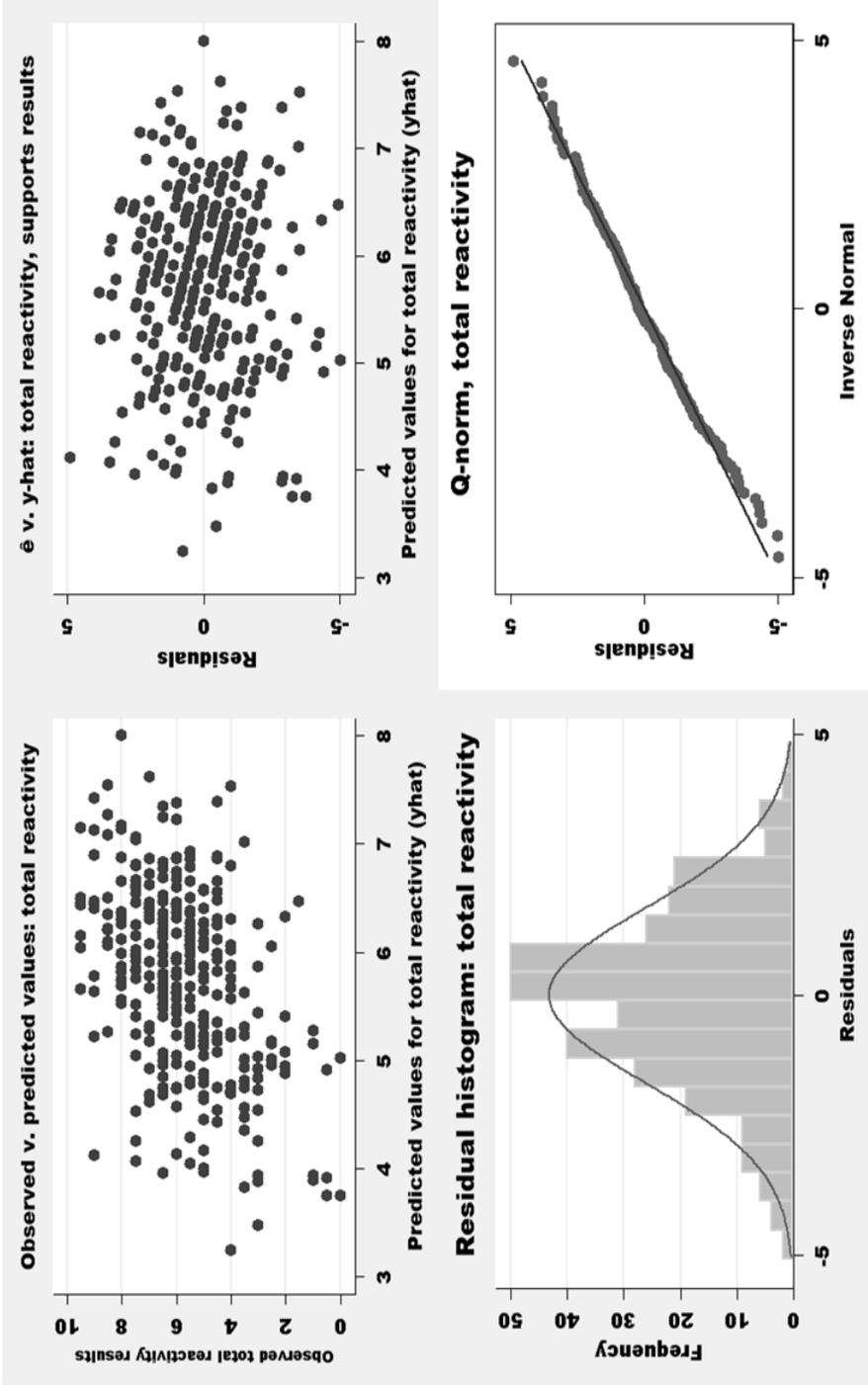


Figure 5.17: Residual analysis for supports data and total reactivity regressions



Incentives to use LSA data

6.1 Introduction

In order to provide professional staff with appropriate motivation to use LSA data in their instruction, it is not enough to write and publish a policy. Changes in practice are difficult, often resisted, and need to pass a 'utility test' to prove their worth to educators with limited time to take on new ways of doing things (especially if the new ways are somewhat unproven in their experience). The right mix between pressure, follow-up and settings reasonable expectations is the key to effective implementation of LSA policy and to ensure the results are used as they are expected to be used in schools (Fullan, 1985; Fullan, 2009; Willms, 2000; Hargreaves et al., 2009).

This chapter is laid out in the following way: (a) a literature review discussing previous work on how different levels and kinds of incentives can affect teachers, as especially how they affect LSAs and data use; (b) a presentation of the findings from the researcher's survey of teachers all across Canada; and (c) conclusions are presented. These results will address several important independent variables: IV16 – expectations; IV17 - follow up; IV18 – results awareness; IV19 - pressure; and IV20, stakes.

6.2 Literature review

The wide-spread use of data in classrooms is dependent, most of all, upon policies that explicitly state that data use is a planned outcome for the assessment program (Klinger, DeLuca & Miller, 2008). Policy alone, of course, does not drive instructional change. The expectation for LSA data to be collected and used originates in the New Public Management ideals from the 1980s (van Thiel & Leeuw, 2002; Fountain, 2002; Kjaer, 2004). Large-scale educational assessments are seen as a way to keep citizens and politicians up to speed on the relative effectiveness of schools, and to provide ideas about how to improve the system (Mintrop & Sunderman, 2009; O'Day, 2002; Nettles & Herrington, 2007).

The list of purposes to which LSAs are intended to be put is also quite lengthy (Benveniste, 2000; Mehrens, 1998; Travis, 1996; Popham, 1999). As noted in the introductory chapter, ministry documents from across Canada reveal at least nine purposes for which tests data can be employed and most provinces tend toward most of these (the province with the lowest number of LSA expectations wants a single instrument to serve five policy functions). So ministries certainly expect the data from LSAs to be informative and instructive.

Expectations for data use must translate further down the educational hierarchy to be effective. School divisions and schools should be on side with the

ministry goals and may possibly even set some goals of their own based on assessment results (Goertz, 2001; Polikoff, 2012). The current model in public schools is to have some form of 'improvement framework' document created at the school level that is checked by the school division and passed on to the ministry. These documents should align with the ministry goals above, and very often use LSA data as the starting point for devising their goals, and also as the key measurement of their success in achieving them. Whether these documents are effective at providing the impetus for instructional change is a part of what this chapter will examine.

Where policies become practice is in the implementation phase. Setting goals is perfectly reasonable, but reasonable people expect that the goals will be pursued and followed-up upon by the leaders responsible for improvement at all levels (Noell, Witt, Gilberton, Ranier & Freeland, 1997; Trouteaud, 2004; Honig, 2004). However, policy implementation is dependent on many factors at each jurisdictional level being aligned and clear to actors in the system. Loss of policy fidelity results when messages are unclear or allow for varying interpretation. The more expectations that are placed on the LSA assessment model, then the more pressing it should be that these expectations are clear all the way down the educational hierarchy and are checked frequently enough to ensure that implementation is being done consistently (Darling-Hammond & Rustique-Forrester, 2005; Witt, Noell, LaFleur & Mortenson, 1997; Elmore, 1980).

High expectations and frequent follow-up are tools to promote the effective collection and use of accurate educational data. They are also a perfect recipe for increased oversight and resulting pressure being applied to teachers who give provincially-mandated tests (Deci, Speigel, Ryan, Koestner & Kaufmann, 1982; Boardman & Woodruff, 2004; Hamilton & Berends, 2006). This pressure is not assuaged in any way by the public scrutiny that follows the release of annual LSA results by ministries which is followed by independent evaluations of school quality from market-oriented think-tanks across Canada (this practice is borrowed from American counterparts). Common twin themes in large-scale testing are the externally imposed pressures of ranking schools, and the internally-imposed pressure that professional educators feel to work as hard as possible to try to get the best results for their students that they can (Espeland & Sauder, 2007; Taylor, Shepard, Kinner and Rosenthal, 2003; Schorr, Firestone & Monfils, 2003).

The higher the stakes and the greater the emphasis placed on standardized tests, the more likely it is that their results will produce reactivity: teachers will change their practices in some way (Espeland et al., 2007).⁸⁵ If sanctions and/or

⁸⁵ Even low stakes tests seem to produce reactivity: "What is striking in our study of teachers' use of assessments is just that - teachers' use. As we have stated elsewhere . . . teachers are using these assessments. Although teachers may not always be using them in

rewards are attached to results, for example, there is a significant impetus to improve raw scores even if there may not be a concurrent impetus to 'generate knowledge' in the abstract sense (Marshall & Drummond, 2006). There appears also to be a direct relation between the level of stakes and the reaction: higher stakes gives educators all the more reason to change practices since the rewards or sanctions are less easily ignored (Cimbricz, 2002).⁸⁶ Policies attaching rewards or sanctions are made at the ministerial or divisional level. The application of intense pressure for results comes in part from the same sources. Yet the actual reaction, positive or negative, occurs with teachers, who have direct access to both students and the tests.

Expecting great results from teachers, "creates incentives to raise test scores per se, not to improve achievement" (Koretz, 2002). Therefore, to avoid the negative pitfalls of higher stakes assessment becomes a real concern. Strengthening a school's capacity to deal well with data demands a team approach with professional learning communities and staff leaders taking the reins (Scott, Webber, Aitken & Lupart, 2011; Volante, 2006). Jurisdictions can make this process easier by stressing the importance of assessment to inform instruction (Scott et al., 2011) and provide essential information about what is working (Shepard et al., 2011). If the educational high road is not taken, the low road will be: test-scores are manipulated too much and too often. Haladyna, Bobbit Nolen & Haas (1991) call the current level of manipulation "staggering" (p.5).

Figure 6.1: Summary of incentives literature

Topic	Author(s)	Summary statement
Expectation / purposes	Ben Jaafar & Anderson, 2007	A comparison of business and ethics-based accountability systems in the Canadian context – they are competing models
	Coburn & Talbert, 2006	Paper on cross-jurisdictional differences in ideas about what is good data and how to use it. Includes lots on score inflation, high stakes, PD, collaboration, implementation, etc.
	Coburn & Turner, 2011	Data use is a construct with a complex web of factors affecting it, thus: jurisdictions, leaders, tools, routines, PD, stakes, etc. are examined.

the way the district intends them to be used, the fact remains that they are consulting, analyzing, and acting on interim assessment results." (Olah, Lawrence & Riggins, 2010, p. 244)

⁸⁶ He goes on to say: "Following this line of argument, high-stakes tests are more likely to impact, if not constrain, teachers' beliefs and practice." (Cimbricz, 2002, p.14)

	Davidson & Frohbieter, 2011	Different ideas on interim LSAs at schools, divisions, classrooms are presented, with information on PD, implementation, and jurisdictional roles.
	Debard & Kubow, 2002	Paper on implementation issues, and the lack of 'bottom up' being a stumbling block of LSA mandates (policymakers need to consult).
	Fountain, 2002	Examines the change in focus from NPM times to a 'customer' model of government services which exacerbates inequalities and non-democratic tendencies.
	Klinger, DeLuca & Miller, 2008	Examination (by province) Canada-wide of grades in which LSAs are done, the purposes of assessments, and historical contexts.
	Kornhaber, 2004	Paper on the testing regime in the US and how test-based accountability is not an answer to concerns nor can it conquer apparent problems in educational system.
	Lachat & Smith, 2009	This paper looks at three urban schools' experiences trying to use LSA data. School leaders are important as are also data analysis skills and PD.
	Mehrens, 1998	Study on cross-purpose testing, test formats, public reactions to results, stakes, students, etc. Ends with reasonable conclusions and some suggestions on how to improve.
	van Thiel & Leeuw, 2002	The performance paradox is the unintended consequences of NPM measures. Paper strong on background and educational examples.
	Wayman, Spring, Lemke & Lehr, 2012	Includes lots of principal strategies to foster data use, but most are not used by study respondents.
	Weinbaum, 2009	Delaware-based 10 state study of AfL enhancement at high schools. The paper focus on high school teams' implementation of AfL vision which shows success only from cases with strong state or school leaders. The rest wallowed and grew little This is basically the story of overworked, under-supported early adopters who didn't change practice as a result of constraints.
	Young & Kim, 2011	A comprehensive literature review on the uses of data. Includes lots of good detailed information and a

		stacked bibliography.
Follow-up	Armstrong & Anthes, 2001	Study of six effective data-use districts that draws general conclusions about school climate and data-use structures.
	Boardman & Woodruff, 2004	Texas study with 20 teachers on the implementation of new reading strategy affected by test pressure/utility.
	Cohen & Ball, 1990	Concludes that teacher implementation is spotty with good reason when policy is incoherent, and built on contradictory or and short-lived plans.
	Desimone, 2002	A review of implementation of CSR showing that implementation in general is strong including some information on jurisdictional roles, PD, and resources.
	Duemer & Mendez-Morse, 2002	Examines policy implementation as qualitative field, and multi-level hierarchies as basis of 'policy mutation.'
	Elmore, 1980	Implementation through backward and forward mapping are spelled out and includes detail on 'street-level' policy discretion.
	Finnigan & Gross, 2007	Chicago study to determine if high stakes increase teacher motivation. The conclusion is positive, motivation and changes in work practices were apparent but, possibly as a result of negative reactivity and support was not forthcoming enough to help teachers with appropriate and effective changes (positive reactivity).
	Halverson & Thomas, 2007	Paper argues that resources teachers (SSTs) are in-house data experts using 2 case studies related to applying this to instruction at school level.
	Harper & Maheady, 1991	Examines the ECRI model teaching implementation and what factors lead to uptake by teachers. It concludes that it works if they is no enforced implementation.
	Honig, 2004	Oakland case study on bottom-up reform notes that reform becomes top-down when policymakers get onto it. Divisional decisions (related to 4 decision-making paradoxes) always favor centralized control.
	Klingner, Boardman & McMaster,	Study on how to scale up evidence-based practices to a large scale. Concludes that scaling up is complex and depends on many actors and factors.

	2013	
	Matland, 1995	Paper comparing top-downers and bottom-uppers in implementation. Includes an interesting matrix of conflict and ambiguity levels.
	Means, Padilla, DeBarger & Bakia, 2009	A US Department of Education commissioned report on how data can and should be used in schools, classrooms, states. Includes 10 case studies in purposively selected districts and has interesting breakdowns of district and school-level supports as well as data interpretation and use criteria for teachers.
	Mintrop, 2003	Maryland and Kentucky case studies in sanctions and how educators respond. They lead to turbulence, denial, and little reflection.
	Thomas, 2007	Looks at the weaknesses of performance measurement, stakes, and game-able accountability systems.
	Trujillo, 2013	A case study on implementing equity policy. States that non-controversial 'top-downs' work, and dispute leads to watering down policy goals.
	Wayman & Stringfield, 2006	Centred on the use of technology to improve use of data, results show some of the benefits of data use and the factors that made it possible.
Stakes and Pressure	Ben Jaafar & Earl, 2008	A cross-jurisdictional comparison of LSA models in Canada from the perspectives of consequences and the use of data. It includes no small-scale data.
	Booher-Jennings, 2005	Examines the practical effects on teachers of high stakes accountability in Texas.
	Cullen & Reback, 2006	Has a lot of statistical data to draw conclusions of results gaming in Texas related to high stakes assessments.
	Ehren & Swanborn, 2012	Dutch schools have evidence of test pool-shaping and not adhering to test administration rules. Examples of negative reactivity shown.
	Haladyna, Bobbit Nolen & Haas, 1991	Looks at test score pollution and how high stakes policies drive poor practices. Includes a chart amended for use in Chapter 3 on ethical/unethical practices.
	Hanushek & Rivkin, 2012	Addresses the use of value-added measurement (VAM) to assess teacher quality and makes note of the effects of measurement error (etc.) but supports the use of VAM.

	Holmstrom & Milgrom, 1991	A paper on an economic model that considers base pay more effective than incentives (especially for teachers in this example). It shows flaws in incentive models in complex equations.
	Jacob, 2004	LSA tests in Chicago are studied where increases appear in one school, but not in another. Evidence of gaming or strategic moves by schools.
	Koretz & Jennings, 2010	Looks at testing policy issues using real world examples of reactions from teachers coaching and the problem of high stakes.
	Propper & Wilson, 2003	This study looks at 3 different cases of public service evaluation. There is an emphasis on 'mis-implementation' of policy and gaming.
	Sahlberg, 2010	An examination of accountability practices and their negative effects. He proposes a different responsibility model.
	Sheldon & Biddle, 1998	Looks at several studies on motivation and concludes reform efforts can go astray with high stakes and sanctions.
	Vernaza, 2009	Florida study using surveys and questionnaire. Some similar purposes to the author's study with some data on reactions to accountability.

6.3 Preliminary hypothesis

H 6-1: Teachers will react to incentives in different ways depending on the amount of pressure they feel is applied to have them improve test scores as compared to improving instruction in general. Thus, an awareness of test scores and follow up from supervisors will increase total reactivity scores with includes both positive and negative effects.

6.4 Results from surveys

The survey results will be presented at the national level, and then province by province looking at the key independent variables in this chapter. The statistical analyses are left to the end, and will look where overall trends can be identified and conclusions might be drawn. These are the variables discussed in this chapter:

Incentives results: These are national and provincial data, and test hypothesis **H 6-1**.

- IV16 (independent variable 16) - expectations
- IV17 - follow up
- IV18 – results awareness
- IV19 – perceived pressure
- and IV20 – perceived stakes

Teachers were asked in an online survey to rate their experiences with the expectations that they use LSA data to improve instruction (from any of three jurisdictional levels) and the follow up on that expectation. Each positive response was scored +1, while the 'none' response for expectation or follow up from any jurisdictional level was scored a -1. The pressure and stakes variables were recorded on Likert scale from 'none' to a 'great deal.' Results awareness was designated a positive score (+1) for any known trend in LSA scores (upward, downward or holding firm) and a negative value (-1) for not seeing the data or not recalling them. Thus ordinal values were converted to cardinal ones based on the discretion of the researcher. Non-respondent data were not used in the analyses.

Note that the chapter-ending charts and tables show each of these variables individually and some further analysis is provided (section 6.8). The values given to survey responses for regression purposes can be seen in **Annex 2**.

6.4.1 National results

- IV16 – expectations (see **Figure 6.7**)

The expectation to use LSA to improve instruction is not strong from any of the three jurisdictional levels that affect teachers' work. The highest proportional rate of expectation from respondents is just over a third (35%) from the school administration which is the top-rated response in five provinces. This drops to 28% for the division level expectations being made explicit (top-rated in three provinces), and to 18% for the ministry (not top-rated in any province). There are also 18% of respondents nationally for whom no expectation was made explicit (rated highest in two provinces).

- IV17 - follow up (see **Figure 6.8**)

The results for follow up are substantially lower than those for expectations, those being not particularly high. The highest response proportionally was no follow up (41% overall and the highest rated response in four provinces) followed by school follow up (39% overall and top-rated in five provinces), division follow up (15% overall) and ministry follow up (5% overall).

- IV18 – results awareness (see **Figures 6.9 – 6.12**)

Results awareness (quite literally being aware of the LSA results) was a much more commonly reported facet of LSAs. Respondents were in 81% of cases aware of class and school results, and in 64% of cases aware of division results. Overall results-awareness was reported at 75% of respondents.

- IV19 – perceived pressure (see **Figures 6.13 and 6.14**)

The pressure reported by respondents was not extreme, but significant. A high amount of pressure affected 38% of respondents, a small amount affected 49%, and no pressure was reported by 13%. It is interesting to note that levels of pressure reported by teachers not giving LSAs were very much comparable. High levels of pressure were reported by a fraction of a percentage more teachers from this group and low levels of pressure were about 10% less frequently indicated. No pressure was correspondingly 10% more frequently reported.

- IV20 – perceived stakes (see **Figures 6.15 and 6.16**)

The level of stakes reported by teachers (made clear in the survey question to refer to those stakes for teachers and schools, not students) follows a somewhat different pattern than that reported for pressure. The highest rated response was medium stakes (44%) followed by low stakes (36%) and then high stakes (20%). With almost twice as much high-level pressure reported as high stakes, this pressure may not come from assessment policy strictures. Pressures reported by teachers not giving LSAs were very similar to those who do give the tests with slightly lower numbers reported for both high and low stakes and therefore more responses in the medium stakes column.

- Summary of incentives data

The data from these questions (all related to policy incentives intended to promote the use of data) pose some problems for analysis. Neither of the initial inquiries into the explicit expectation to use data or the overt follow up on that expectation elicited a strong response. The question related to stakes also seems to show that built-in policy incentives are not considered excessive, motivating or threatening.

My thing with it is that it is not the be-all, end-all. At the same time it is not something . . . I mean I would never say it is okay *not* to do the FSA, it's okay *not* to do provincial exams. They need to happen, but, if they went away tomorrow would I be upset? No. I think there is way more time and stress and energy put into them than what we get out of them. - **BC, Division staff, male**

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

No [expectations to use data from the division level], none. None whatsoever. We choose to do [common assessments] because that's the only way we can grow as a school: monitor the progress of our students. And it is important for us to have that data. . . to try to become a more data-driven school and show how, umm, how we can improve as a result of using the data from previous years.

- NB, High school principal, male

Absent these factors, we are then left with the understanding that teachers are well aware of their students' results (and in particular their own students) and that the pressure teachers feel to use the data and to help students achieve is strong. Interviews with teachers will bear out that while policy incentives do not drive behaviours, it appears that the intrinsic motivation of educators to do their level best for their students does seem to have a large effect. Hargreaves et al. (2009) speak of this kind of motivation as intrinsic and having more value than extrinsic accountability motives.

Yeah. . . even though there hasn't been evidence of that [expectation to use the data], I still think when you say to a teacher the words provincial assessment or standardized assessment, I mean I think they still have that at the back of their minds, right?

- MB, Elementary school principal, female

Teachers, I think, try and do the best job they can. I think they really, truly want to do a good job. So that's of where I kind of start from. Some people might take a different approach and say, 'If you don't get on someone they are going to be lazy or try to take the easy way.'

- ON, High school principal, male

Some teachers just have a professional, umm, ethic, work ethic, that they, they really, umm, they constantly delve into what they can do to do better. - PEI, K-9 school principal, female

The teachers knew that these test scores were going to reported back to the staff or the school division and that there was some, I don't know, there was some pride, I guess, that we wanted to do well and that we wanted to not let our kids down if they were going to be assessed in these areas. So teachers respond to that.

- SK, Elementary school principal, male (a)

As provincial results are examined, interview data from province-specific teachers will be shown to support this early conclusion.

6.4.2 Provincial results

Alberta

- IV16 – expectations (see **Figures 6.7 and 6.9**)

Alberta has a fairly typical grouping for data use expectations. The highest rated response is school-level (42%) with a declining trend across all other responses in order: division-level (29%), ministry-level (17%), and no expectation (12%). Only the 'school-level expectation' and 'no expectation' responses vary much from the national norms, the school-level being 7% higher and no expectation 7% lower.

- IV17 - follow up (see **Figures 6.8 and 6.9**)

Follow up in Alberta does not deviate too far from national norms. School-level follow up is somewhat higher than the national average (53% to 39%) and no follow up is somewhat lower 35% to 41%). Both division- and ministry-level follow up scores are slightly lower than those recorded nationally.

- IV18 – results awareness (see **Figures 6.10 – 6.13**)

Alberta teachers are higher than the national average in total results-awareness with 80% of teachers reporting being aware of class, school, and division data (the national average is 75%). The awareness is most pronounced for classes (89%) and declines at the school (84%) and division (68%) levels.

- IV19 – perceived pressure (see **Figures 6.14 and 6.15**)

In terms of perceived pressure, Alberta teachers are again very close to national norms. The highest rated response was 'a small amount' (53% compared to the national average of 49%). The second most common response was 'a great deal' with 38% of respondents (nationally this figure is also 38%). Only 9% reported feeling no pressure.

- IV20 – perceived stakes (see **Figures 6.16 and 6.17**)

Reported perceptions of stakes differ from pressure significantly, and also differ from the national ratings. Most respondents (40% compared to 44% nationally) reported a medium level of stakes. With the highest rating nationally, 33% of Alberta teachers report tests are high stakes for teachers (nationally this is reported at 20%). Low stakes was reported 9% less commonly than in the national average (27% to 36%).

- Summary of incentives data

Alberta teachers report feeling somewhat higher expectations to use results and somewhat more follow up on this expectation than is true nationally. This is a factor that leads to the high rating for stakes in Alberta where more teachers consider LSAs high stakes for teachers and schools than in any other province. The

assessment program touches more teachers here than in most provinces (the ratings for 'no expectations' and 'no follow up' are quite low).

The current administration is content with the fact that we are talking with each other and we talk to them and we discuss the concerns that we have. Everything tends to be a proactive approach. I think that's one of the values I share with my current administration. . . I think our administration trusts our judgment on this one. Again, that's at a school-level. From the board-level I'm not sure they think that.

- **AB, High school Science teacher, male**

The teachers here are not much more aware of their results (except at the class-level) than teachers elsewhere, and this overall awareness metric only just tops the national average by 5%. The difference lies mostly in their awareness of class-level results, placing them second in the country. This focus on class-level results is a factor of the pressure that teachers feel regarding provincial testing.

The pressure is there from administrators, from the district, from wherever to increase performance without necessarily talking about what performance means in the context of these subjects, just looking at the averages.

- **AB, High school Science teacher, male**

I understand why the government feels that we need that kind of accountability, but I feel grade 3 is too young to write that many tests. Five tests in four weeks is a lot for an eight-year-old to take on. . . Unfortunately, I know that it is meant to improve teaching practice but I think it actually does the opposite because people feel the pressure and feel pressure to teach to the test. - **AB, Elementary English teacher, female**

It can be seen in Alberta that teachers' average positive reactivity rating is the highest in the country, so it might be assumed that higher expectations for data use and follow up coupled with teachers feeling a higher amount of pressure to get good results creates the necessary conditions for positive reactivity. This may well be the case (and it will be examined in the analysis that follows), but negative reactivity does in the end outweigh the positive in Alberta (by an average of 1 rating point), so the necessary conditions are in place here for both kinds of reactivity effects.

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

British Columbia

- IV16 – expectations

British Columbia teachers reported overwhelmingly that there is very little expectation to use data from any jurisdictional level. Fully 86% report no expectations (the national average is 19%), while school (8%), division (3%) and ministry (3%) ratings are all well below the national averages.

- IV17 - follow up

Again at the far extreme of provincial ratings, BC teachers reported in 94% of cases there is no follow up on instructional change. The national figure for this metric is 41%. Only school- and ministry-level follow up even rated (both at 3%) for follow up being done.

- IV18 – results awareness

Teachers in this province are middling regarding results-awareness. Overall 67% of teachers were aware of some LSA results. Teachers here are also just about as aware of school-level data (70%) and division-level data (64%) as they are of student-level results (67%). BC teachers rate lowest in Canada for class results-awareness, but in the middle of the pack for other data in this section and fall only 8% below the national average for overall results awareness.

- IV19 – perceived pressure

BC teachers had a very equal distribution related to the question of pressure. Most respondents noted a small amount of pressure (56%) while both the great deal of pressure and no pressure responses were at 22% of responses. Only Nova Scotia indicates a similar near-equal proportional split between high and no pressure responses. Nationally the high pressure response (38%) is higher rated than the no pressure response (13%).

- IV20 – perceived stakes

Teachers rated stakes for teachers and schools as quite low as British Columbia had the lowest ratings of perceived stakes nationally. LSAs were rated as low stakes for 61% of respondents, 28% rated them as medium stakes, and only 11% rated them as high stakes. Compared to the national data, the low pressure response in BC is highest in Canada and the medium stakes response (at 44% nationally) was the lowest figure in the nation.

- Summary of incentives data

British Columbia teachers have provided an interesting and atypical set of responses to questions in this chapter. They report both the expectation to use data

and the follow up on instructional change as barely evident in their practices. The ratings for these factors were the lowest in Canada. Yet the data analysis that follows shows BC teachers are just about as aware of results, except at the individual student-level, as most teachers nationwide. Respondents were unusually inclined to dismiss the pressure they felt (with the second-lowest national rating) and the stakes involved (the lowest rating).

It can be concluded that in the absence of expectations to use the data or follow up on these missing expectations, teachers are still aware of the results data.

Partly is was the math, that year when I showed the staff the [poor] results, that made them think, 'Oh maybe we have to do something different' but it wasn't 'Oh gosh, we have to do it.'

- BC, Elementary homeroom teacher, female

I don't think it is appropriate for a teacher 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

Since British Columbia has the lowest national figure for positive reactivity, awareness may not translate into effective and appropriate use of the data. Of course, the ratings for negative and total reactivity are also quite low, so results-awareness does not necessarily lead to reactivity at all.

My teaching partners continuously change, all the grade four teachers. I've been there for many years. I think that in the beginning I felt some pressure. I think maybe the newer teacher might feel a bit, but I just alleviate with saying, 'You know what - we're not going to worry about these results'. . . There are so many that are not getting the support they need that they are not going to do well on the test anyways. We can't let it get to us, but I think it probably does, but we try not to let it.

- BC, Elementary homeroom teacher, female

It is like any other piece of information. And my understanding of data is that you want at least three pieces of information when you are making a decision, if possible. So if you've got a bunch of classroom stuff that says the kid is doing fine; you got a bunch of

stuff from home, the former teacher, etc. and they happen to bomb a provincial exam or an FSA, well maybe it is not a big deal.

- BC, Division staff, male

It appears that in the assessment system in British Columbia the expectations for data use do not get passed on to teachers very well despite the admonitions of the provincially ministry.⁸⁷ They also do not appear to be very keenly aware of or highly reactive to the pressure or stakes written into provincial LSA policy as compared to other provinces.

Manitoba

- IV16 – expectations

Manitoba teachers reported very low levels of expectations to use LSA data in their practices. 64% reported no expectation, and at all three jurisdictional levels the proportion of responses was well below national norms. Only British Columbia has a similar distribution of responses.

- IV17 - follow up

Again Manitoba teachers report similar responses to those from BC teachers in that the follow up reported was very low. A low 69% report no follow up and all jurisdictional levels are below their respective national averages.

- IV18 – results awareness

In terms of this metric, Manitoba has the least results-aware teachers in Canada. It has been noted (in the introduction) that the LSAs employed in elementary and middle years grades are quite different in Manitoba and this is likely a reason for this divergence from the national data. A teacher-rated checklist may be good data for a teacher to examine, but they are not particularly good for inter-school or inter-division comparisons.

- IV19 – perceived pressure

Manitoba teachers had more respondents indicate no pressure than any other jurisdiction. The proportion who reported a small amount is 5% lower than the national average, and high levels of pressure were 14% less common than is true nationally. This may again be a result of the nature of the testing done in Manitoba which is unique in Canada.

⁸⁷ "The BC Performance Standards are intended as a resource to support ongoing instruction and assessment." British Columbia Education, retrieved Aug. 9, 2014 from: http://www.bced.gov.bc.ca/assessment/fsa/pdfs/fsabrochure_print.pdf

- IV20 – perceived stakes

The stakes that teachers report in Manitoba are quite low and in line with the other metrics from this chapter. Low stakes were reported by 50% (the national figure is 36%) and only 18% reported high stakes (nationally this figure is 20%).

- Summary of incentives data

These data paint a picture of assessment policy in Manitoba that differs from national norms in all measures that are done. Manitoba also differs from other provinces in that the responses from this province are all near or outside the bounds of national averages. The low rating for expectations, follow up, results-awareness and both pressure and stakes puts Manitoba at the extreme edge of this set of data – these aspects of assessment policy are not well known or well-regarded here.

Currently I would say that [motivation to use data] is just personally [applied]. Back when it first started I think it came from our principal. . . And as far as the division goes, I really don't know the last time this was brought up at the school division. Like the focus has moved so far away from there with focusing on new report cards and focusing on different things that I don't even know the last time this was brought up and mentioned. - **MB, Middle years Math teacher, female**

I wouldn't say about the standard test, the grade three and the middle years one, I wouldn't say [conversations] are common. I would say we do them, we have some conversations around them, usually the teachers who administered it, myself and our resource teacher, you know. Nothing ever shocking comes out of it. . . The provincial one is a piece of [school data-based decisions], but it is certainly not a big part.

- **MB, Elementary school principal, female**

The provincial ministry has set out clearly the goals for their assessment program (as in the introduction), but these data do not indicate that the goals are either understood or complied with by in-service staff.⁸⁸ Manitoba is also among the least reactive provinces nationally (see Chapter 3) and these data may point to some reasons as to why this is the case. As has been often repeated throughout this

⁸⁸ "The Provincial Assessment Program supports learning by: providing feedback to students, teachers and parents about student learning; informing instructional planning and helping to determine the need for changes or student specific interventions..." Manitoba Education, retrieved Aug. 9, 2014 from: <http://www.edu.gov.mb.ca/k12/assess/>

section, a large determinant of the reactivity may well be the nature of the assessments that are done here.

So I know there were some discussions saying like, 'Well our division goal is that 80% of our kids have a 4 and we are nowhere near that, so what is the problem?'. . . There was not a lot of discussion on how the data were collected or what was going on. So they say, 'It needs to be fixed,' but then when we were pushing for ideas on how we could make it more consistent so we could actually see where the students were falling they weren't interested in that. They would say that it is up to your professional judgment.

- **MB, Middle years Math teacher, female**

So I've had experience with high school, with some, with the standard [-ized] test. . . And the anxiety around it, and the teachers feeling the anxiety as well. You know, I don't know that I, like I just don't agree with it. - **MB, Elementary school principal, female**

What is clear is that the strictures of other jurisdictions' assessment policies are not evident in Manitoba and as a result pressure is decreased, stakes are more manageable, but reactivity tends toward the negative (overall) and is not strong compared to other provinces.

New Brunswick

- IV16 – expectations (see **Figures 6.7 and 6.9**)

New Brunswick teachers report expectations very much in line with national scores. School expectations are highest rated (46%), followed by divisional expectations (28%), ministry expectations (19%) and finally no expectation (8%).

- IV17 - follow up (see **Figures 6.8 and 6.9**)

The responses to questions about follow up are again close to national norms in New Brunswick, and thus help to uncover the trend that expectations are not always aligned with a proportional follow up. School follow up was strong (48%) but was trailed by the no follow up response (27%) and only then by division (21%) and ministry (3%) responses. The same trend is nationally apparent.

- IV18 – results awareness (see **Figures 6.10 – 6.13**)

Overall results awareness is quite high in New Brunswick (82% compared to 75% nationally). Teachers are most aware of classroom data (88%) followed by school data (86%) and then division data (71%). As results-awareness is relatively high, the results-awareness for all three jurisdictional levels was rated higher than was true nationally.

- IV19 – perceived pressure (see **Figures 6.14 and 6.15**)

Pressure reported by New Brunswick teachers is quite substantially higher than is true for the national data. The 'great deal of pressure' and 'small amount of pressure' responses were equally rated (at 48% of responses) leaving the 'no pressure response' at a mere 4% of respondents. There are only three other provinces with high levels of pressure at the same or greater levels than the low levels response, and all of these have very low response numbers for 'no pressure.'

- IV20 – perceived stakes (see **Figures 6.16 and 6.17**)

Teachers in New Brunswick rate the stakes for teachers and schools as somewhat more than we see in the national data, but still not extreme. The highest proportional response was for a middling level of stakes (50%) followed by low stakes (28% compared to 36% nationally) and high stakes (22% compared to 20% nationally).

- Summary of incentives data

It has been noted (in Chapter 3) that reactivity in New Brunswick is strong in the positive category (2nd highest nationally), middling for negative (5th nationally) and the net effect tends slightly to the negative. From the data in this section it can be seen that both pressure and stakes are more prevalent in this province than is true nationally, results-awareness is higher than national figures, and both the expectations for use and follow up are slightly more apparent. The independent variables here all point in the same direction: that there are strong centralized and local expectations to use LSA data.

The survey data put New Brunswick near the lead of provinces with clear and enforced expectations policies, but interview respondents indicated that there were some uncertainties for teachers regarding the expectations for the data.

A teacher is a teacher because of their intrinsic drive. I don't think they need any external forces pushing them to be better.

- NB, Middle years homeroom teacher, female

Data related to results distribution were analyzed in Chapter 4, yet respondents also made clear that the data themselves did not always get distributed in such a way that they were made sufficiently results-aware to take action. Results-awareness would certainly suffer if the presentation of results was haphazard, non-uniform, or not done. There are examples of data presentation and clear expectations being made explicit.

It is district-driven. We have a school improvement plan that we have to do every year and basically it is instructional-based. . . So we do that at the beginning of the year. When we go in in August we 're gonna talk about what are we looking at that we really want to focus

our instruction on. So using the data, the district requires that we fill out an SIP. - **NB, Middle years homeroom teacher, female**

Everything here is done through formal assessment strategies, right? So they have to have, umm, and we have common exams, so common exit exams. So they spend a lot of time working on that.
- **NB, High school principal, male**

The LSAs in New Brunswick are intended to support school-level decision making and teachers in their instructional choices.⁸⁹ So while there is a good alignment in all the independent variables considered in this chapter, there is not the kind of reactivity effects expected if incentives and explicit expectations alone were sufficient to drive instructional changes based on the results data.

Newfoundland and Labrador

- IV16 – expectations

Expectations for data use in Newfoundland and Labrador are almost item-by-item identical to those from New Brunswick (as seen above). They also show a slightly higher proportion of school-level expectations than the national figures (45% to 35%) and a smaller proportion of no expectations (10% to 19%).

- IV17 - follow up

Responses related to follow up on instructional change in this province are quite similar to national figures. Follow up is also strongest from the school-level (48% compared to 39% nationally) but nil at the ministry-level. No follow up is reported by 38% of respondents while the national figure is 41%.

- IV18 – results awareness

Overall results awareness is virtually identical with national data. 75.2% of teachers reported awareness in Newfoundland and Labrador compared with 75.0% nationally. Teachers are 7% more aware of school results and 7% less aware of class results.

⁸⁹ “. . . [LSA data] enables teachers to . . . determine the needs of their students, and to address those needs adequately in order to tailor instruction. . . [LSA data] enables policy makers to make programming decisions at the . . . school level.” New Brunswick Department of Education and Early Childhood Development, retrieved Aug. 9, 2014 from:
<http://www.gnb.ca/0000/results/pdf/AssessmentFrameworkDocument.pdf>

- IV19 – perceived pressure

Newfoundland and Labrador teachers reported the highest proportion of respondents feeling 'a great deal of pressure.' The figures in this category are 16% higher than the national average (54% to 34%). Lower levels of pressure were reported by fewer teachers, but it is telling that the no pressure response was chosen by only 8% of respondents compared to 13% nationally.

- IV20 – perceived stakes

The responses regarding stakes were much closer to being in line with national scores. Medium levels of stakes were the most common choice (50%), and 23% report high stakes for teachers (20% nationally), and 27% report low stakes, compared to 36% nationally. Stakes are somewhat less prevalent in this province than reported across Canada.

- Summary of incentives data

One reason that Newfoundland and Labrador figures for expectations were compared with New Brunswick numbers at the beginning of this section is to highlight again the apparent difficulty in assigning strong correlations to the factors examined in these chapters, at least in isolation. This province, like New Brunswick exceeds the national averages on most ratings considered here: higher for expectations, higher for follow up; about equal for results awareness; much higher for pressure; and only lower slightly for stakes. New Brunswick results are much the same as these. Considering the vast difference in reactivity effects found in these two jurisdictions (see Chapter 3) it is hard to believe that the policy factors considered here could have much of an overall relationship to reactivity.

It is certainly true that Newfoundland and Labrador teachers feel a significant amount of pressure to have their students do well on LSAs:

Before I write the public exam . . . we have three weeks, of every single day going to teach them how to . . . manage multiple choice tests and every single item in the acids and bases is to be covered. And then after seeing five or six of the same types of questions, they're ready for it. And then they'll know exactly what they are expecting. So, their marks are high, everybody's happy, they're getting their scholarships, they get their entrance marks.

- NL, High school Science teacher, male

But this (often) self-administered pressure does not equate to high stakes tests for the teachers – there was no evidence reported to the researcher of poor test results leading to sanctions or serious consequences for any educator. High stakes test for students (as those administered here at the high school level) may be high pressure for teachers, but they do not appear to promote positive reactivity as much as the negative. Newfoundland and Labrador teachers have one of the most

negative national ratings for net reactivity (in front of only Ontario and BC), but they are also third highest in total reactivity. Prioritizing positive effects over the negative would reduce the extremity of both of these ratings, and in so doing come closer to meeting the positive pedagogical goals teachers themselves express about their practice.

If the exam were removed and I could do what I liked, well then there would be a hell of a lot of pressure removed. You know I feel under a lot of pressure, but you know, it's self-induced, really, to make sure they do well, and the students are under that same pressure. . . You see across the hallway there are students over there that do courses that don't have provincial exams and they seem to really be enjoying their courses . . . But in my room, you know, I don't want anyone knocking on the door from 9:00 until 10:00 - were doing problems and nobody moves.

- NL, High school Science teacher, male

Nova Scotia

- IV16 – expectations

Nova Scotia teachers reported expectations to use data in line with national figures. Over four categorized responses, these data are never more than 5% away from the national average and follow the same declining pattern from schools, to divisions, ministries, and then no expectation.

- IV17 - follow up

The follow up on instructional change is slightly less apparent in Nova Scotia than nationally. School follow up figures are nearly identical, but the highest rated response is no follow up (45%, which is 4% higher than the national average). The small difference comes from a matching decrease in figures for division and ministry follow up.

- IV18 – results awareness

Overall results-awareness is reported by 63% of teachers and this is 12% less than the national average for this metric and the second lowest nationally. Nova Scotia teachers are less aware of class results than is nationally true (76% to 81%), less aware of school results (67% to 81%) and division results (44% to 64%), with the lowest national rating here.

- IV19 – perceived pressure

There is less pressure reported by Nova Scotia respondents than in the nation-wide sample. The 'small amount' of pressure response is very near the national average (48% to 49%) but the higher pressure response is less prevalent (25% to 38% nationally) and the no pressure response is more evident (27% to 13%).

- IV20 – perceived stakes

In line with the pressure responses, teachers here report low stakes as well. In fact, aside from British Columbia the stakes reported here are the lowest in the nation. Of all respondents who give LSAs, 5% report high stakes, 39% report medium stakes, and 56% report low stakes.

- Summary of incentives data

The data from Nova Scotia regarding incentives is not what the author expected to see considering the reactivity data covered in Chapter 3. It is a common argument in the United States that assessment policies that build in expectations, incentives and consequences are the best way to ensure compliance with their intentions, and this same argument is not unheard of in Canada. It has been noted several times already that the only province with net positive reactivity is Nova Scotia (if only by a small fraction) and now it can be concluded that expectations and incentives here are generally only as high as national averages and in many cases less evident than is true nationally. There are certainly variations between schools and school divisions in this regard.

There is an understanding that we use all available assessments to guide instruction; this is emphasized in Education classes and PD consistently. There is no direct pressure from administration to use the results in a prescribed way.

- Anonymous survey comment

Oh yes, that [expectation to use data] is very explicit. . . And I think sometimes that is what certainly causes teachers angst and stress as well, too. I think that it takes a lot of work. . . It takes a lot of practice, a lot of focus, you know, a lot of planning. And I think some teachers find it stressful, certainly. But they really, it has been, there has been a big push to do that [provincial assessment] in the system.

- NS, Division staff, female

Perhaps the conclusion being already imagined at this point should be not about positive reactivity, but about total reactivity, and then these variables begin to make more sense. Nova Scotia is the least totally reactive province, and this could be the result of poorly designed incentives and expectations in the policy itself. The stated objectives of the Program of Learning Assessment for Nova Scotia (PLANS; see the introductory chapter) include collecting and sharing information for all stakeholders from students and parents all the way up to ministry officials, but we should also consider the reported low levels of stakes for teachers seen here. Perhaps it is more telling (since less common) to consider the reported low levels of pressure (only Manitoba teachers report more 'no pressure' responses).

These data seem to display very little motivation to use LSA results to improve instruction in classrooms. Certainly this is not the only use to which these data can be put, but using them this way (as PLANS has itself laid out) may be the best way for Nova Scotia teachers to employ the results in helping Nova Scotia students to meet curriculum outcomes.

Ontario

- IV16 – expectations (see **Figures 6.7 and 6.9**)

Expectations to make use of LSA data are clear to teachers from the school level (8% above the national average at 43%), the division level and the ministry. There are also fewer no expectations responses here than is true in the national sample (9% to 19%).

- IV17 - follow up (see **Figures 6.8 and 6.9**)

Follow up on expectations are reported to be equal to or above national norms across all three jurisdictional levels. School-level follow up is strongest (52%), but division-level (20%) and ministry-level (5%) follow up is also reported. No follow up is proportionally quite low as a result, at 23% compared to 41% nationally.

- IV18 – results awareness (see **Figures 6.10 – 6.13**)

Ontario is the second highest rated province for overall results awareness at 85% (the national figure is 75%). In contrast with some other provinces, the class-level data are not as clear to teachers as school results and division results. These large-scope data are more readily used for comparisons of schools and divisions rather than instructional change.

- IV19 – perceived pressure (see **Figures 6.14 and 6.15**)

Ontario teachers report high levels of pressure, in some respects, the highest nationally. There are no respondents who felt no pressure. Some provinces have somewhat higher proportions of teachers who feel high levels of pressure, but no province show the same 100% response to feeling some amount of pressure. The high pressure response in Ontario is also higher than the national average (45% to 38%).

- IV20 – perceived stakes (see **Figures 6.16 and 6.17**)

Ontario reports the highest national levels of stakes as well. The medium stakes response shows the highest proportion in Canada at 57% of respondents. The high stakes response was rated at 29% of teachers, 9% higher than the national average and tied with Québec for the second-highest rating nationally. Low stakes, not surprisingly is 22% lower here than the national figure.

- Summary of incentives data

Ontario and Nova Scotia fall next to each other alphabetically and serve therefore as excellent contrasts in looking at the effects of incentives on reactivity. Nova Scotia shows low amounts of policy incentives to encourage data use,

whereas Ontario shows very high levels of incentives. Nova Scotia also shows net positive reactivity; however, Ontario has the highest net negative score in Canada. Nova Scotia is the least reactive province overall; this contrasts with Ontario which is much more reactive and fifth nationally in this regard.

Incentives to use data promote the use of LSA results in schools in Ontario:

I don't think there is [much pressure] except for English teachers and for math [the two LSA tested subjects] . . . they feel there is a lot of pressure. The math teachers because it is reported as the math results for that teacher, I think that they feel like they are being compared, which they are... the English teachers, I feel that they have a personal investment in the success of their students. But myself, as a science teacher? No. I could completely ignore the tests. Someone else administers it. It has no bearing on my day-to-day work.

- **ON, High school English consultant, female**

It is pretty hard sitting around the superintendent level of table if you've got to cover for your schools that frankly suck in comparison to your colleagues. You know, those are difficult conversations. And what is it about what you're doing with your schools that allows for them to be performing at this level and my schools. . . they might have the same access to the same PD, well why is it we are not performing? - **ON, Division staff, male**

Ontario also has an odd mixture of elementary and middle years assessments (which are not unlike those across the nation), but also a high school-level minimum competency exam in literacy which is not nearly as common as the grade 12 exit exams (which usually are high stakes for students in terms of grades). Even without the high stakes exit exams, teachers in Ontario report very high levels of pressure and stakes applied to them. This is at least in part because LSA results are used to compare teacher performance (as well as school and districts).

In elementary, yes, absolutely, because it is very easy to track that information. In secondary it is a little more difficult - the math results I know absolutely are used to analyze teachers' performance although that information has never been released. I've heard Superintendents and union people say that it is a comparison that is done between teachers. I don't think anything comes of those results, but it is absolutely used.

- **ON, High school English consultant, female**

Absolutely... absolutely, and central board staff who would be in charge of literacy and numeracy, they would do extensive studies on the results. - **ON, High school English consultant, female**

When we sit in our principal's meetings and . . . our superintendent, he shares everyone's data. So there is nowhere to hide. So he'll throw pie graphs up there, he'll throw bar graphs up there saying okay. . . Here are all the grade 9 compulsory courses in your building. This is what the success rates were in terms of pass/fail for all these courses. So, you know, when you look at some schools versus others, you know, as a principal you kind of sit there and say, 'Man, I can't believe, you know, 10% more kids failed this class in my school than they did somewhere else.' . . . So I mentioned EQAO results. . . for the schools that come up sometimes at a 65% success rate where the provincial average is 85 [%], yeah. Is there pressure for them to do better? For sure. Absolutely. - **ON, High school principal, male**

It is true that Ontario's Education Quality and Accountability Office (EQAO) has done their best to educate not only teachers but also the public about the dangers of using LSA results to draw conclusions about schools or even worse, individual teachers. Yet perceived pressure and stakes remain high. Some of the pressure is self-applied as many teachers reported doing everything they could to see their students achieve good results but most often in situations where expectations were made clear and school administration were supportive.

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

So, the average classroom teacher there, no, I don't think there was much expectation that they use the results in their daily practice. I think it was we're either doing really, really, really well and we should rest on our laurels, and keep doing what we are doing, or we are not doing well, but we haven't done well for 3 or 4 years so there's not much hope.

- **ON, High school English consultant, female**

If I have to do this [as a teacher] . . . and I know it is expected and I know it is important, but just tell me what I need to do and explain it to me clearly and concisely so that I can convey it to the kids. I think

when you get into lesson plans that allow too much freedom, especially when you are doing something that is just so clear and concise, it muddies it for the teacher, and by extension, the kids won't get it clear in their heads either.

- ON, High school principal, male

In the final analysis, though, the tilt toward negative reactivity effects appears to make clear that policy incentives are not the clearest route to positive, curriculum-rooted, high-quality instructional change.

Prince Edward Island

- IV16 – expectations

The expectation to use data is seen by respondents from Prince Edward Island to be somewhat greater than the national average and differs from the national data in that school, district and ministry expectation ratings all fall within 2% of a 30% rating. This is the most equitable distribution of expectations nationwide. The 'no expectation' response was less frequently chosen (9%) than was the case nationally (19%).

- IV17 - follow up

PEI teachers also rated the follow up on these expectations as being more evident than the national averages indicate. The 'no follow up' response was proportionally lower than the national average (34% to 41%). There is slightly less school follow up than the national figures (34% to 39%), but again the main difference here is the prominent role played by the ministry. Division and ministry follow up were rated equally at 16% of respondents which is the highest rating for ministry follow up nationwide, and the only province where the ministry takes as active a role as the division (the national averages for the division and ministry are 15% and 5% respectively).

- IV18 – results awareness

Respondents from PEI had the greatest level of results awareness in the country rated at 86% (the national average is 75%). Teachers had the highest level of class results awareness (94%), the second highest level of school results awareness (94%) and the third highest level of division results awareness (70%).

- IV19 – perceived pressure

The perceived pressure reported by teachers from this province was also close to the highest levels across Canada. Compared to 38% nationally, 53% of PEI teachers reported high levels, 44% report lower levels and 3% report none.

- IV20 – perceived stakes

Stakes were also highly rated in PEI with high stakes and medium stakes responses well above the national averages (high stakes - 27% to 20%, medium

stakes – 53% to 44%). The low stakes response was correspondingly below the national average (21% to 36%).

- Summary of incentives data

The expectations data from Prince Edward Island are similar to those from Ontario, and here we are able to look at an 'apples versus apples' comparison (the previous Nova Scotia / Ontario comparison was 'apples versus oranges'). All five metrics in this section were rated higher than national averages by these two provinces. Each of these provinces also had the nationally highest ratings in at least one category. Yet when looking back at our Chapter 3 reactivity data, the connection begins to fray. PEI teachers reported net reactivity just barely in the negative, and the lowest rate of net negative in the nation (Nova Scotia is net positive). Ontario is the most net negative province. PEI is well down the list for total reactivity (6th) and Ontario is in about the same position (5th nationally). PEI rates third in positive reactivity effects (which is the major reason they rate as high as they do in both total and net effects) and 7th for negative effects. Ontario rates only 8th for positive effects and 1st for negative effects. There is a virtual flip-flop of positioning relative to the kind of reactivity that is most prevalent in each province.

These seemingly contradictory data do not make it easy to understand how incentives policies affect reactivity. With virtually identical incentives ratings from Ontario and Prince Edward Island, their use of positive and negative reactivity practices could scarcely be more different. It is certainly true that PEI teachers are aware of the incentives and of externally applied pressures to perform.

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

- PEI, Elementary homeroom teacher, female

PEI didn't do very well in the PISA testing around mathematics so there was a lot of public outcry that our students aren't doing well. You know we finished in think 9 or 10 out of 10 of the provinces. . . [We need] to change the curriculum, and need to do this, and we need to add days to the school year. . .' So I'm starting to think that public opinion is starting to push that expectation that our students do better in mathematics. **- PEI, High school administrator, male**

I know our grade 9 Math teacher feel incredible stress. . . We talk about the journey being a ten year journey and so if you're missing some essential foundational learnings, and umm, then you have a

math assessment at the end of grade 9 and you haven't really achieved, I think our grade 9 teacher feels like, you know, a lot of pressure for her. . . She thinks it looks bad on her. We keep trying to remind her that they didn't '*not learn*' in over one year, they '*not learned*' it over a number of years.

- PEI, K-9 school principal, female

The striking difference between PEI and all other provinces examined thus far is the large part the ministry plays in setting expectations and making them explicit to teachers as well as the role they have in following up on these same expectations. It is easy to imagine that the ministry role here is not as isolating as it might be in some other jurisdictions. Charlottetown, home of the ministry, is less than a two hour drive from the most distant points on the island and much closer to most. Compared to any other province, this puts the ministry in close quarters with schools. Having this close physical proximity makes it possible for the message to retain some fidelity (Davidson & Frohbieter, 2011; Fullan & Pofret, 1977; Shower, 2010) that it might otherwise lose in the translation from the ministry to the division, and then to the schools in larger provinces. The rating is still quite low, but by the measures of this survey, PEI has the most hands-on ministry nationally.

I think that they expect the provincial results in addition to the school-based results to inform our instruction, to, umm, look at our goals for the schools, for, I guess future planning for what our schools are. **- PEI, K-9 school vice principal, female**

We actually have, umm, you know, our team that are [*sic*] responsible for our school effectiveness, whenever we do set out our goals. When we go to our director, the director looks at the goal and says, 'Okay, what are we basing this on?' Then we have to go back to the data from the assessments and based upon what we see as an area of need, that is what the goal has to reflect.

- PEI, K-9 school principal, male

I would think first and in most cases it would be like an internal pressure. But from school to school, you know some of the administrators might handle things a little bit differently. They might have a, more of a hands-on approach and teachers might feel pressure from school administration. When it comes to the board and the department [or ministry]. . . I don't think people would feel pressure from those levels. **- PEI, High school administrator, male**

Québec

- IV16 – expectations

The expectations for teachers to use LSA data are at a similar level but different in their application in Québec as compared to the nation as a whole. School expectations are almost identical to the national average (35%), but division expectations are higher (37% compared to 28% nationally). Ministry expectations are lower (9% to 18% nationally) and the 'no expectation' response was about the same at 19%.

- IV17 - follow up

Respondents rated follow up just about identically to the national figures (all four responses are within 4 % points) and they show the same pattern for most common response down to the least common.

- IV18 – results awareness

Québec teachers also are just about as results-aware as the national sample. Overall results awareness and school results awareness are just about matching, yet teachers are more aware of class results (86% to 81%) and less aware of division results (59% to 64%) by relatively small margins in each case.

- IV19 – perceived pressure

Teachers report slightly higher levels of pressure than the national sample. The high pressure and low pressure responses are inverted compared to the national sample. There is more high pressure in Québec (50% to 38%) and less low pressure (39% to 50%). The no pressure rating is just 2% lower than the national figure.

- IV20 – perceived stakes

Stakes are also rated as more being evident in Québec than the nation. While the most common responses were low and medium stakes (both rated at 36% of respondents), high stakes were reported by 29% of teachers which is higher than the national figure (20%).

- Summary of incentives data

The pattern in Québec related to incentives for teachers to use LSA data is not clear at first glance. Expectations for their use are high (and higher from the division level than is usual), and both stakes and pressure are reported as higher than the national norm. Yet teachers are no more aware of the results from the LSAs than jurisdictions where expectations, stakes and pressure are not as highly rated. It is also reported that for all the expectations that are laid out, follow up is not as prominent. There seems to be some concern that expectations can be unrealistic, and that follow up is not evident according to interviews respondents.

So they had looked at the results in some sort of PLC at the board, and they had decided what our goal was going to be. This 3% [increase] from the 75% average to a 78% average [LSA score], which

is arbitrary and ridiculous. Anyway, [we] said how, 'do you have some suggestions as to how we can do that?' Well she said 'you could collaborate with each other and you know, share information.' Well, we collaborate with each other constantly. . . 'Oh [she said], well you could try to do common assessments during the year.' We do a common assessment at least once a year. . . Absolutely, we do that. 'Well, um. Well, um. I don't know.' . . . And understand that that person is a teacher that was made a consultant. So when you switch from this chair to sitting in another chair it doesn't give you any special abilities or anything else.

- **QC, High school English teacher, male**

I think [pressure] just comes from having that exam. I guess from the ministry. Yeah, just from knowing that is coming up, you know, even weighing my decision about whether or not I wanted to loop [the practice of teachers staying with students for consecutive years] with my kids or not. I wanted to loop but I knew that that [the exam] was at the end of the year. I was going to have to face that exam, those exams. - **QC, Middle years homeroom teacher, female**

Québec teachers self-reported more negative reactivity than the positive sort (see Chapter 3) and total reactivity is quite high (2nd highest in the country). One might be drawn to conclude that the missing factors in Québec policy incentives (follow up and results-awareness) might just be the reason why differentiation between positive and negative practices is not made. It might also be surmised that the present policy factors (expectations, stakes and pressure) promote reactivity, but do not promote one type of reactivity over any other. Interview subjects confirmed that there was some disconnect between policies and practices in this province.

There is a tendency in Québec for high school teachers in particular, and even elementary teachers, to work on their own. . . There is one of the teachers in our department who is very weak. It is not looked upon as our [the department's] problem. It is the administration's problem and it is of course the students' problem, and it *is* our problem in the sense that when we get these students come into our classes now we have to rectify the situation. . . Nobody would ever dream of speaking to this teacher and saying, 'Look, you need to pull your socks up. Can we help you? Can we work together? Can someone mentor you?' - **QC, High school English teacher, male**

Saskatchewan

- IV16 – expectations

The expectations for teachers to use LSA data to inform instruction in Saskatchewan are similar to the national average. The only noticeable divergence from the national data is the fact that divisional expectations (34% to 28% nationally) are proportionally higher than school expectations (29% to 35% nationally). Neither ministry expectations nor 'no expectations' responses were more than 3% from national sample data.

- IV17 - follow up

The follow up reported by teachers here was slightly greater than the proportions reported by the national sample. That said, the greater number of divisional follow up responses (23% to 15% nationally) makes up about the entire shortfall in no expectations responses (32% to 41% nationally).

- IV18 – results awareness

Overall results awareness is less common in the Saskatchewan sample (68% to 75% across Canada). Teachers here are less aware of results across all measured jurisdictions: 2% less class results-aware, 14% less school results-aware, and 7% less division results-aware.

- IV19 – perceived pressure

The pressure reported by teachers in Saskatchewan is not dissimilar from the amounts reported by the national sample, but it is reported to be present in different ways. There is less high pressure reported (33% to 38%) and more low level pressure reported (57% to 49%). There is also 3% fewer no pressure responses from the provincial sample as compared to national data.

- IV20 – perceived stakes

The perceived stakes for teachers and schools is substantially less than the national average according to Saskatchewan teachers. Only 10% report high stakes (20% in the national data), and this 10% difference is made up equally between a 5% higher proportional rating for medium stakes and 5% more for low stakes.

- Summary of incentives data

By most measures Saskatchewan teachers report expectations proportions that are similar to or less than what is true for the national sample. The only category where teachers report a higher response proportion is regarding follow up on instructional change. Although this data does ask teachers to recall the Assessment for Learning (AfL) LSA tests which have not been administered over the last 2 school years, it seems clear that recollections are relatively vivid since the responses are not dissimilar from the national data. The only category where teachers report a higher response proportion is regarding follow up on instructional change, but it should be noted that this was not true for all respondents.

I don't think that expectation [to use data] was ever given to us on what to do with them. It was sort of like, I felt like, oh, 'You have to give it to see where students are,' and then there was no follow up from anyone. It was just, this is what it is . . . and that was it.

- SK, Middle years homeroom teacher, female

[The expectation to use data] was regular in terms that it was more just 'What are you doing about it?' There was that emphasis that you'd better be doing something about it. They weren't really helpful in terms of *what* you are doing, it was more like, 'Are you doing something about it? You better be.' . . . It was more a pain in the ass than anything. Sorry. . . In fairness to them, what could they offer? I mean, I guess their job was to make sure it was still up front and that I needed to keep on the right path.

- SK, Elementary school principal, male (a)

I don't think for my personal teaching that me giving out an AfL is going to make the next kid read better or to do better in school. So I didn't, I never used them for marks or never used them. And I knew before I ever gave it to a student how that student, I probably could have gauged pretty well, how that student would do on it before ever giving it to them.

- SK, Middle years homeroom teacher, female

It is also noticeable that the divisional role in Saskatchewan LSAs is a prominent one. The expectations to use LSA data was more from the division level than for the school level here, and that is true in only two other jurisdictions. This expectation, whether a result of its origins or not, does not translate well into awareness of the assessment results. In this regard Saskatchewan teachers trail the national data by a substantial margin.

You sometimes get, not just related to that data, I think sometimes you get a group of students who are, as a whole, very strong academically, and then sometimes you'll get a group of students who aren't. And that sort of follows them through but based on one AfL I wouldn't say it is indicative of how it would go throughout [their schooling].

- SK, Middle years homeroom teacher, female

The high levels reported for the divisional role is something of a riddle here. The role of the ministry is not highly rated, and respondents regularly rated

the division higher than was true in the national data. Interviews indicated that the divisions often served as the connection between assessment policies and practices. With many small schools in evidence in this part of Canada, perhaps this is a necessary role in cases where teaching staffs are not always large enough to be expert in assessment. It should be noted that divisional input was in some cases (but not in all cases) construed as productive.

Every year as administrators we kind of review and modify our [division-mandated] goals. It is good to have, the more data that you have, it makes that process more authentic.

- **SK, Elementary school principal, male (b)**

Like, I'll give you one example. Our school division was excited because we were above the provincial average in math and we scored 2 out of 5 in a certain category. And at a staff meeting I just said, 'Why are we celebrating this? Like our kids scored 2 out of 5. That is 40%. We flunked.' Like, so what if we were above the provincial average. . . Shouldn't we be trying for 5 out of 5? . . . Maybe understand why our kids only got 2 out of 5? [Instead] we kind of got patted on the back because we were above the division average and above the provincial average.

- **SK, Elementary school principal, male (a)**

6.5 Correlation analysis – incentives

This matrix uses Spearman's rank order correlation techniques to examine the inter-relatedness of independent variables and provincial variations regarding these same factors. Relationships are designated as weak ($p < 0.05$) or strong ($p < 0.01$) and can have positive or negative signs attached to them depending on whether there is agreement or divergence in the variables.

The independent variables in this chapter are once again closely, positively, and significantly correlated (see **Table 6.2**). The highest levels of correlation are between expectations for data use and follow up (0.472) and between perceived pressure and perceived stakes (0.401). As a means of securing some use of the data, expectations and follow up appear to apply a sufficient amount of pressure. It should be noted that what is classified as high stakes in Canada is most likely seen as rather weak stuff by the observers from other countries. Still, there are too many documented cases of cheating and deception for anyone in any jurisdiction to take for granted a simple dynamic between high pressure and good professional conduct (Rhoades & Madaus, 2003; Jacob & Levitt, 2003; Simner, 2000; Amrein, Berliner & Rideau, 2010). Noting here that stakes and

pressure are also significantly and positively correlated to results awareness, it seems that these various incentives variables are quite well aligned in the sense that the values assigned by the researcher to survey responses point uniformly in the same direction.

Figure 6.2: Spearman's rank order correlation test done with incentives variables

Correlation matrix - incentives variables

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

* p<0.05; ** p<0.01

6.6 OLS regressions – incentives

6.6.1 Regression analysis

Reactivity is the dependent variable of this study and (see Chapter 3) has both positive and negative effects. There are also two further aspects of reactivity: total reactivity which examines all reactivity, positive and negative, and net reactivity which determines whether positive or negative effects are more prevalent. Net reactivity is not examined below since it is premised on cancelled out values, and thus gives an incomplete picture of the data. The other reactivity options will be shown and discussed in this order: positive reactivity; negative reactivity; and total reactivity. Provincial dummy variables were added to examine variation at the jurisdictional level, but please note that MB is the control province for positive reactivity (it does not have dummy added), BC for negative reactivity, and PEI for total reactivity.

The **positive reactivity** data (see **Table 6.3**) show that there are three significant factors from this chapter on incentives. The strongest correlation is between **perceived pressure** and positive reactivity, and this is significant at the p<0.01 level both before and after provincial dummies are added to the regressions. The two other independent variables are significant only at the p< 0.05 level both before and after the inclusion of the dummies, and have weaker correlations. These are the **follow up** on data use and the **perceived stakes**.

Table 6.3: Positive reactivity is correlated against independent incentives variables.

Positive reactivity

Expectation to use data	0.062 (1.40)	0.032 (0.69)
Follow up on data use	0.120 (2.26)*	0.127 (2.40)*
Results-awareness	0.039 (1.38)	0.044 (1.58)
Perceived pressure	0.778 (3.99)**	0.764 (3.94)**
Perceived stakes	0.356 (2.06)*	0.374 (2.16)*
(provincial dummies) AB		0.290 (1.14)
BC		-0.456 (1.76)
NB		0.094 (0.37)
NL		-0.024 (0.08)
NS		0.061 (0.23)
ON		-0.010 (0.04)
PEI		-0.592 (2.31)*
QC		0.093 (0.33)
SK		-0.490 (1.61)
Constant	1.755 (13.37)**	1.870 (8.93)**
R²	0.18	0.24
N	343	343

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

The data on perceived pressure bears out what has been evident from interviews from the beginning: that teachers feel a significant amount of pressure regarding LSAs done at the provincial level. Some of this pressure was based on community or parental expectations. Some came from a desire to not be singled out by published rankings. The greatest driver of this pressure was reported to be 'self-imposed.' Most teachers stated that having the professional drive to do their jobs well meant they wanted to see their students achieve at high levels. In interviews it became clear that the external factors were less important to teachers than this intrinsic motivation. There can be no doubt that pressures internal and external do have a significant and strong impact on positive reactivity practices.

The next strongest correlation for an independent variable in this section is the follow up on data use. This was reported by teachers to occur somewhat sporadically, but where it was happening, it was most commonly from school-level administration rather than divisional or ministry staff. This factor is certainly an aspect of LSA policy implementation that could use some further examination. Implementation literature is fairly clear that expectations and policy goals are not sufficient to inspire change – there must be some follow up and accountability for actors in the system (Witt, Noell, LaFleur & Mortenson, 1997; Noell, Witt, Gilberton, Ranier & Freeland, 1997; Garn, 1999; Cohen & Ball, 1990). The fact that it was so sporadically reported shows that this has not been fully factored into Canadian educational policies.

The final significant variable is also the weakest relationship examined in this section. Respondents reported the level of stakes they felt was applied to teachers and schools by LSA policy. Generally levels reported in the survey were not exceptionally high (especially when compared to American examples where teachers can be fired and schools can be closed), but they do seem to have a significant if small impact on the use of positive reactivity strategies.

It stands to reason that if a teacher feels they have more to lose they will be motivated by loss aversion to avoid any sanction. Yet the low level of stakes reported does align with what interview subjects had to say since no respondents knew of any case where a teacher was fired or re-assigned based on LSA scores. That does not mean that teachers do not perceive stakes being high, but they are also limited by the realities visible in schools and codified in their contracts.

Altogether these variables account for 18% of the variance in reactivity scores prior to the addition of provincial dummies. This is quite a high proportion and seems to indicate that specific incentives are an important aspect of LSA policy to ensure effective and proper implementation to achieve ministry goals.

Looking at variations between provinces, there is only one weakly significant negative correlation and it is with Prince Edward Island data. They therefore diverge from the control group (MB) to show less positive reactivity.

Table 6.4: Negative reactivity is correlated against independent incentives variables. **It is important to note that since negative reactivity is enumerated in negative integers, a negative coefficient means more negative reactivity effects, not less.**

Negative reactivity

Expectation to use data	0.056 (1.11)	0.012 (0.23)
Follow up on data use	-0.002 (0.03)	0.019 (0.32)
Results-awareness	-0.070 (2.18)*	-0.044 (1.43)
Perceived pressure	-0.787 (3.58)**	-0.652 (3.05)**
Perceived stakes	-0.207 (1.05)	-0.019 (0.10)
(provincial dummies) AB		-0.271 (0.98)
MB		0.304 (1.08)
NB		0.058 (0.21)
NL		-0.225 (0.72)
NS		0.978 (3.43)**
ON		0.457 (1.53)
PEI		-0.310 (1.10)
QC		-0.235 (0.78)
SK		0.676 (2.03)*
Constant	-2.511 (17.18)**	-2.795 (12.06)**
R²	0.07	0.18
N	344	344

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

Overall, PEI is the third most positively reactive province overall, so this divergence from control group seems to indicate that incentives factors in particular are less effective in this jurisdiction. If provincial dummies are included in the regressions the amount of variance explained by the variables increases to 24%.

There are two incentives variables that have significant impacts on **negative reactivity** prior to the inclusion of provincial dummies, and both of these relationships are themselves negative (see **Table 6.4**). Teachers who reported feeling **pressure** as a result of large-scale provincial testing were more likely to use negative reactivity strategies. This finding follows closely upon what has been seen in terms of positive reactivity; that perceived pressure is a key factor in reactivity practices, positive and negative.

A weaker negative correlation between **results-awareness** and negative reactivity appears in these data, but this variable was not rated as significant in terms of positive reactivity. This finding demonstrates that teachers who are more aware of their class, school and divisional results are somewhat more likely to employ negative reactivity strategies. Since negative reactivity tends to favour practices that do not depend on a close analysis of the data (they are not particularly data-informed, even if they do 'work'), this may indicate a facile evaluation of the results, counting on the test having common content year on year, or 'light-lifting' attempts to improve scores. This finding also aligns well with a finding from Chapter 4 which showed that detailed returned data (aggregated and disaggregated scores) has a significant correlation to more reactivity. Since the relationship does not appear after provincial dummies are added, it should be considered with some caution.

Before provincial data are included, these variables have an R^2 value of just 7%. Nova Scotia has a strong and significant positive correlation with the control group (BC) indicating less negative reactivity. Saskatchewan has a smaller but still significant positive correlation which also indicates less negative reactivity. These two provinces have the lowest rates of negative reactivity responses in the nation, so it stands to reason that they would diverge in this direction from the norm.

Adding the absolute values of positive and negative reactivity scores provides a picture of the total amount of teacher reactivity related to LSAs (see **Table 6.5**). The regression for **total reactivity** effects shows that the variables highlighted for total reactivity are closely related to those seen in the negative reactivity table while the positive reactivity variables with weak correlations are missing (follow up and perceived stakes).

Table 6.5: Total reactivity is correlated against independent incentives variables.

Total reactivity

Expectation to use data	0.007 (0.09)	0.034 (0.43)
Follow up on data use	0.118 (1.30)	0.103 (1.14)
Results-awareness	0.123 (2.53)*	0.099 (2.07)*
Perceived pressure	1.622 (4.91)**	1.489 (4.53)**
Perceived stakes	0.536 (1.82)	0.363 (1.23)
(provincial dummies) AB		0.845 (2.23)*
BC		-0.129 (0.30)
MB		0.060 (0.14)
NB		0.371 (1.00)
NL		0.482 (1.10)
NS		-0.697 (1.69)
ON		-0.195 (0.49)
QC		0.618 (1.44)
SK		-0.865 (1.81)
Constant	4.225 (19.01)**	4.310 (12.07)**
R²	0.17	0.23
N	336	336

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

That is, the total reactivity effects of these two variables show comparable t statistic values to those which appear in the positive and negative tables – the effects are reinforcing. The preliminary hypothesis of this chapter stated that both positive and negative reactivity would be increased with additional incentives to use the data, and this appears to be the case for perceived pressure.

Perceived pressure has a strong, positive, and significant link to total reactivity, and **results-awareness** has a somewhat weaker positive significant correlation. As these variables have been just discussed, it will be left only to say that the cumulative effect of these variables on total reactivity is quite large prior to the addition of provincial dummies.

A relatively large 17% of the variance in reactivity responses can be explained by these variables. Despite the fact that the relevant independent variables remain identical and their relative levels of significance do not change the R^2 value only increases when we add the provincial dummies to the regression, up to 23%. Provincial variation goes some way to further explaining these variances.

Only one province shows a significant correlation regarding these variables and the control group. Alberta has a weakly significant positive correlation indicating that there is somewhat more total reactivity in this province in comparison with the control group (from PEI in this case).

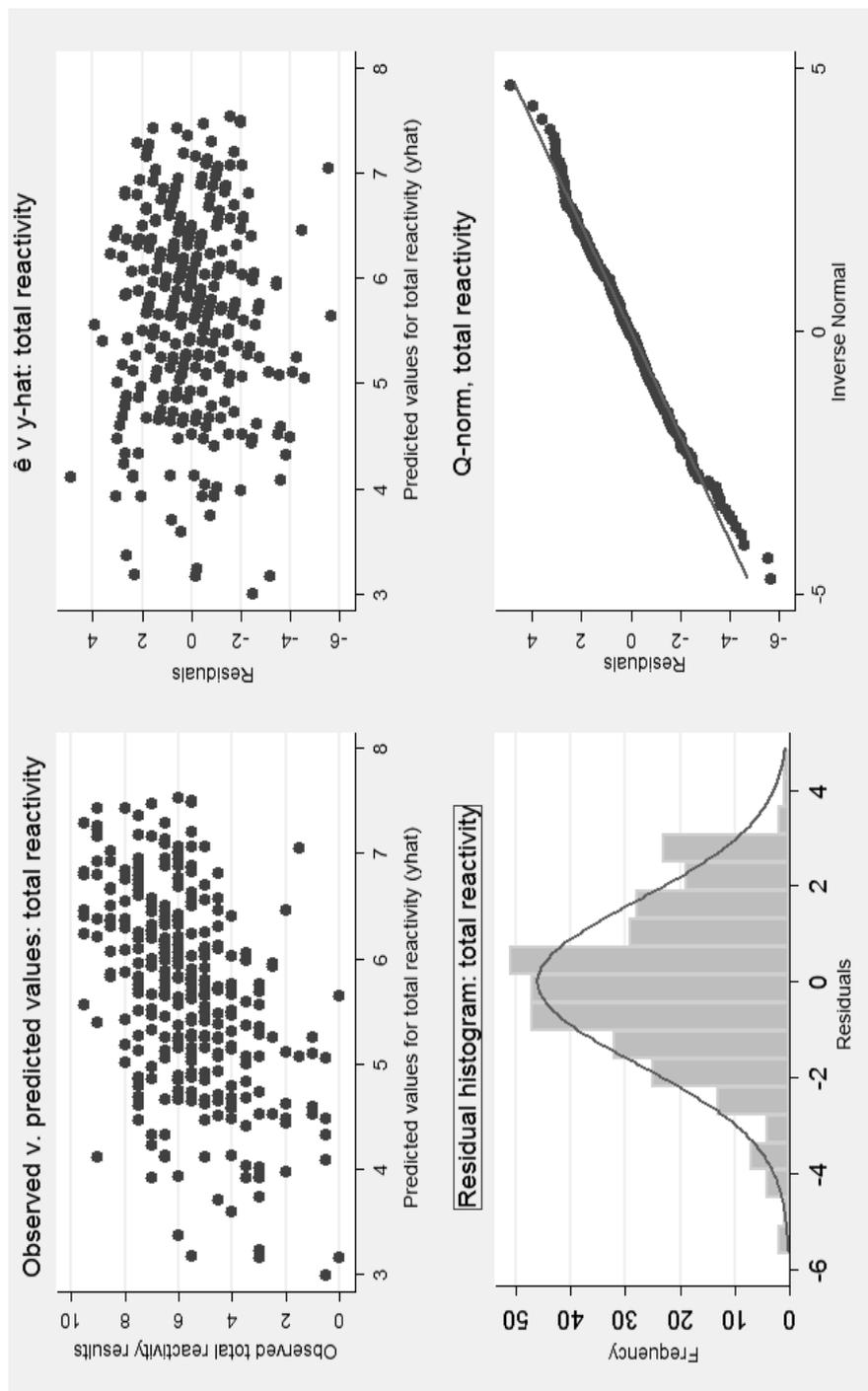
In sum, the regressions here show that if there is one single factor that drives reactivity to results data, it is perceived pressure. The downside of this finding is that the incentives do not serve to have teachers discriminate between positive and negative effects, so that this pressure inspires teachers to work closely with the data to improve instruction and broaden outcomes for students (positive effects) and also to use the data to seek those strategies suited to improving scores on these tests, but that have little leverage beyond them (negative reactivity).

6.6.2 Residual analysis

The residuals from these regressions were examined using four different econometric graphing techniques and the results from these analyses were fairly uniform across all three regressions. The results for the **total reactivity** residual examinations are found in **Figure 6.6**. Total reactivity includes both the positive and negative reactivity results, so it is in some ways a more comprehensive tool than either of them on their own.

The 'observed v. predicted values' chart shows a weak linear trend. Having no visible linear trend would challenge the significance of the regressions. The bands seen in the graph indicate different reported levels of reactivity (from 0 to 10 by multiples of 0.5).

Figure 6.6: Residual analysis for incentives data and total reactivity regressions



The ' $\hat{\epsilon}$ v. \hat{y} ' chart has these same bands, but no clustering and only small numbers of close outliers. More serious clustering or the presence of extreme outliers would make the regression coefficients less accurate.

The residual histogram has a relatively normal distribution, and so the residuals are more likely to meet the assumption of normality and independently distributed residuals required for hypothesis testing using OLS methods when they are this close to normality.

Finally, the QQ plot shows only minor deviations from the normal distribution in the tails. In all, these analyses bear out the rigour of the regression model and help to confirm the findings. (**Figures 6.18 and 6.19** show the other two residual analyses from this chapter.)

6.7 Conclusions

The initial hypothesis in this chapter asserted that both positive and negative reactivity would increase based on incentives variables.⁹⁰ This theory bears the scrutiny of the facts just presented. The amount of pressure that teachers perceive is the most telling correlation from these data and it does impact reactivity just as was earlier surmised. Pressure, and results awareness to a lesser extent, serves as a catalyst to teachers using reactivity strategies, positive and negative.

There is voluminous literature that examines the unintended consequences of high stakes and highly incentivized testing. Very little of this body of research supports the educational value of such practices. These data support the sceptical perspective that incentivized testing does not necessarily promote positive educational practices or positive improvements in instructional strategies. Only with specific and ongoing training, support and professional development (as seen in Chapter 5) can teachers and school leaders expect to guide reactivity toward the positive path. It should be said that even though divisional supports proved most effective at guiding teachers in this direction, instructional leadership does not always come from those drawing higher salaries – it often comes from peers who are also in front of students and as a result whose opinions are respected (distributed leadership in assessment is discussed in Noonan & Renihan, 2006). The school-based sharing of data has also been shown to be a key driver of positive improvements in teaching and this staff culture depends on having 'a community of trust' which cannot be built sustainably upon a foundation of pressure. In

⁹⁰ **H 6-1:** Teachers will react to incentives in different ways depending on the amount of pressure they feel is applied have them improve test scores as compared to improving instruction in general. Thus, an awareness of test scores and follow up from supervisors will increase total reactivity scores, which includes both positive and negative effects.

conclusion, while pressure clearly 'works' to increase reactivity effects, this statement needs many caveats to highlight its downsides.

6.8 Charts and tables

Figure 6.7: Provincial and national data on the perceived expectation for teachers to use LSA data.

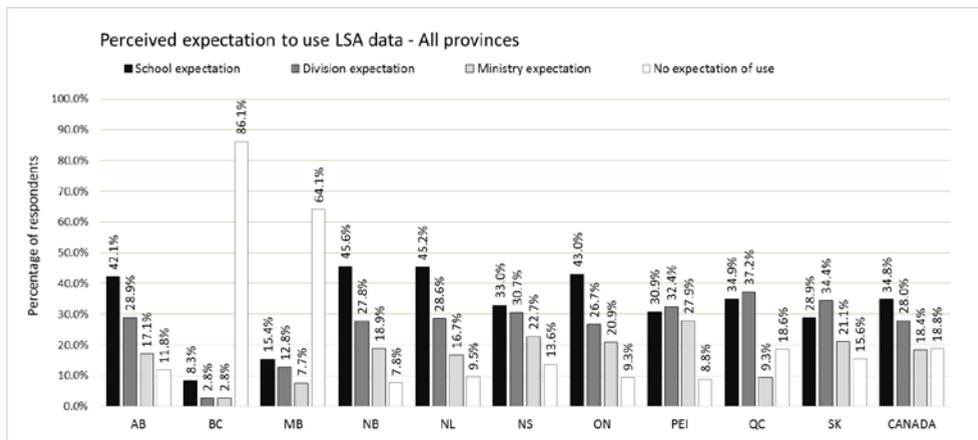


Figure 6.8: Respondents rated how common instructional change was followed up upon.

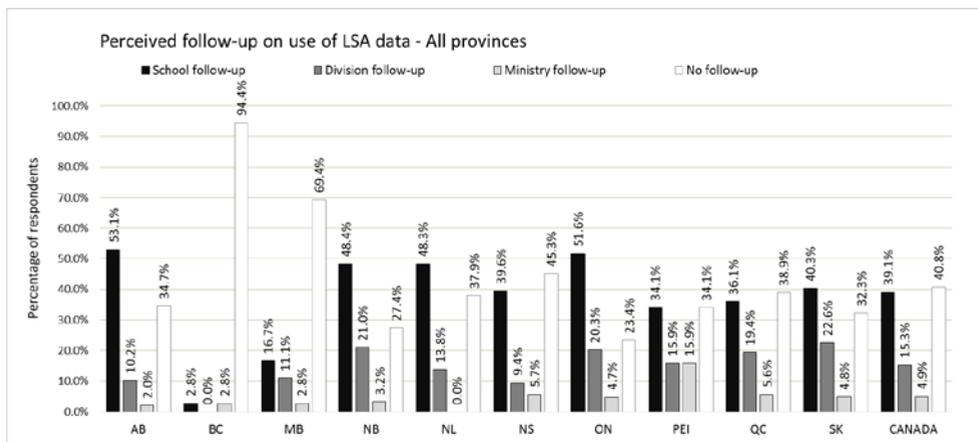


Figure 6.9: Breaking down expectations and follow up on data use by jurisdictions.

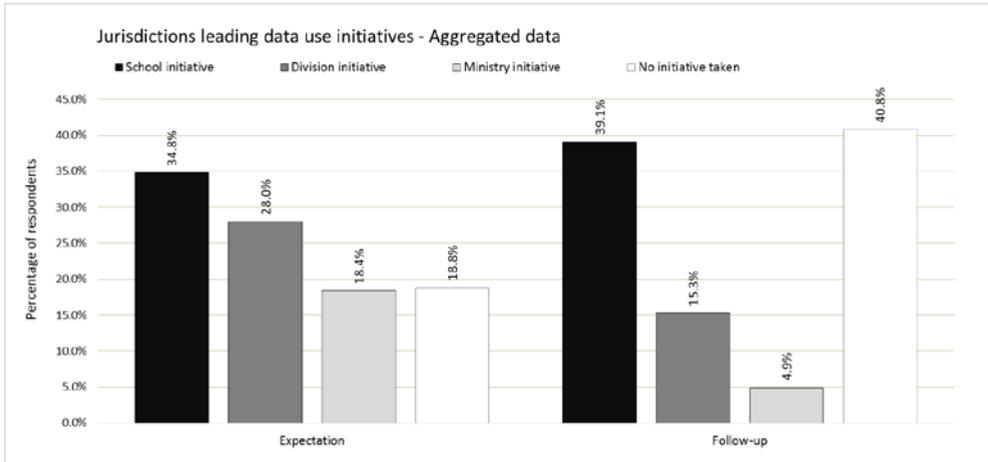


Figure 6.10: Teachers rated how aware they are of class-level data.

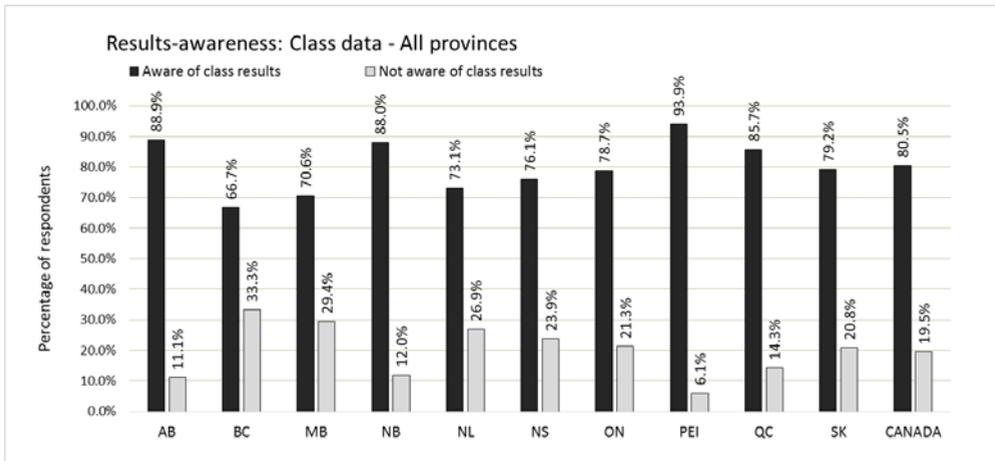


Figure 6.11: A rating of respondents' awareness of school-level data.

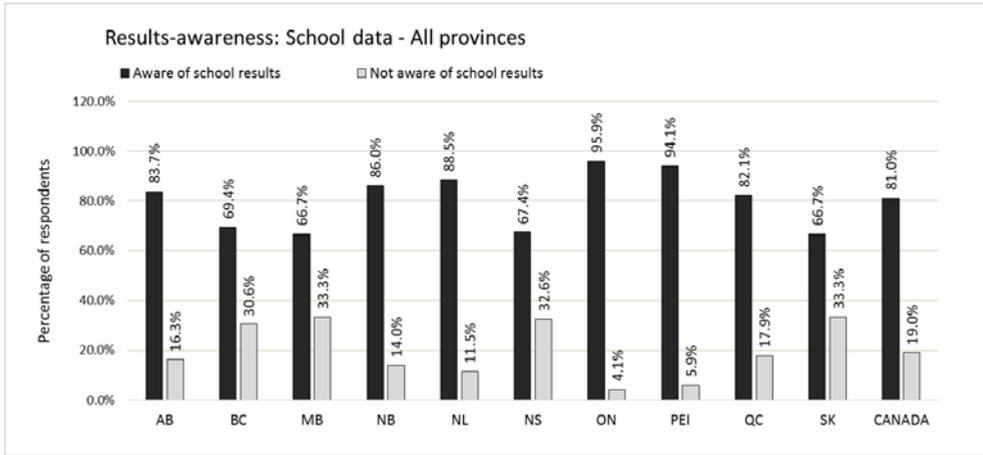


Figure 6.12: Awareness of division-level data was also self-rated by teachers.

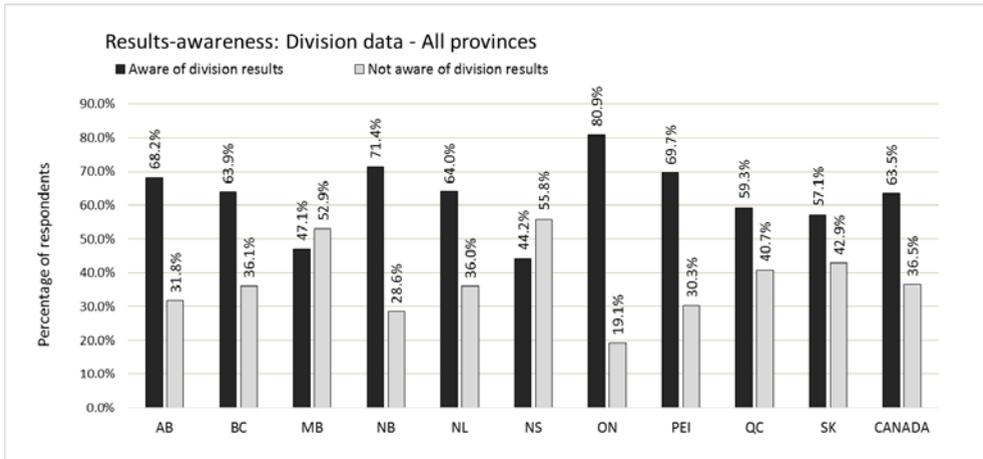


Figure 6.13: Averaging awareness scores across class, school and divisional levels.

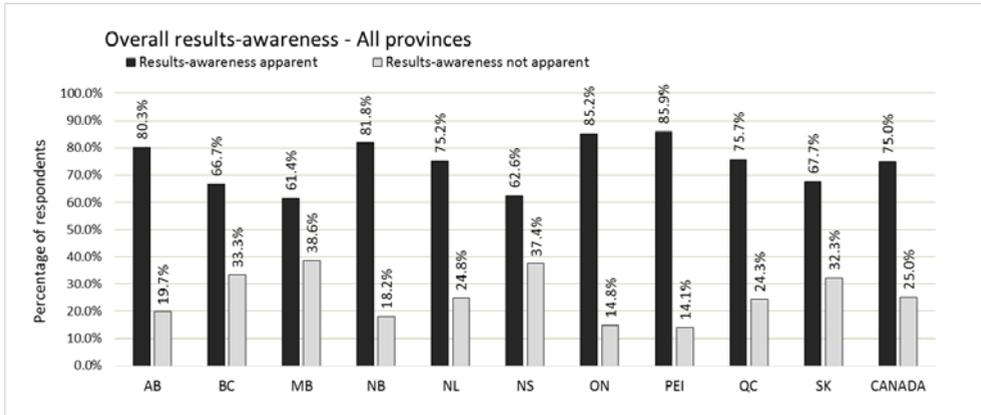


Figure 6.14: Teachers reported how much pressure they feel in relation to LSA testing.

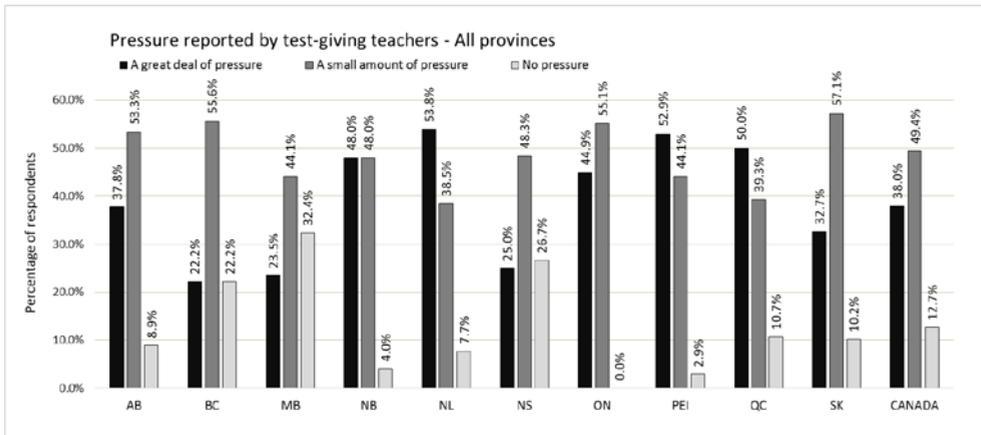


Figure 6.15: Teachers who do and teachers who do not give LSA tests are compared in terms of perceived pressure.

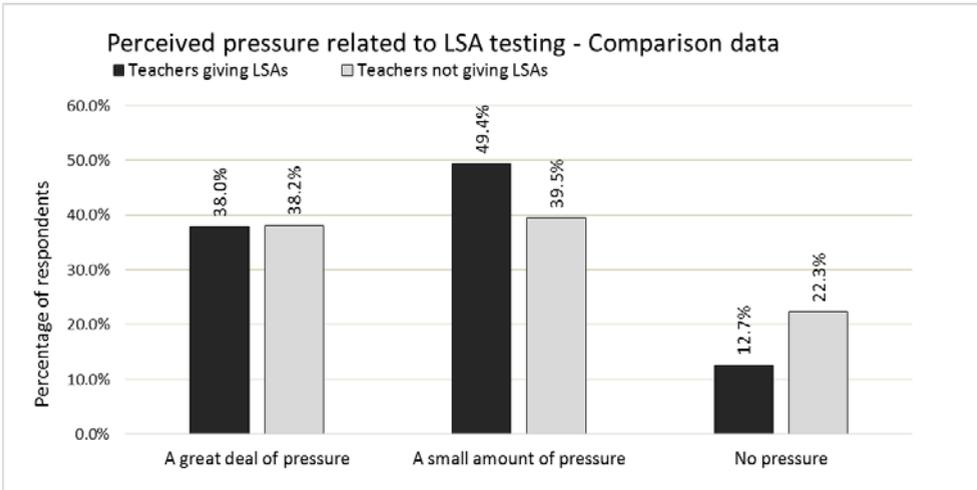


Figure 6.16: Teachers rated the level of stakes they think are applied to teachers by LSAs.

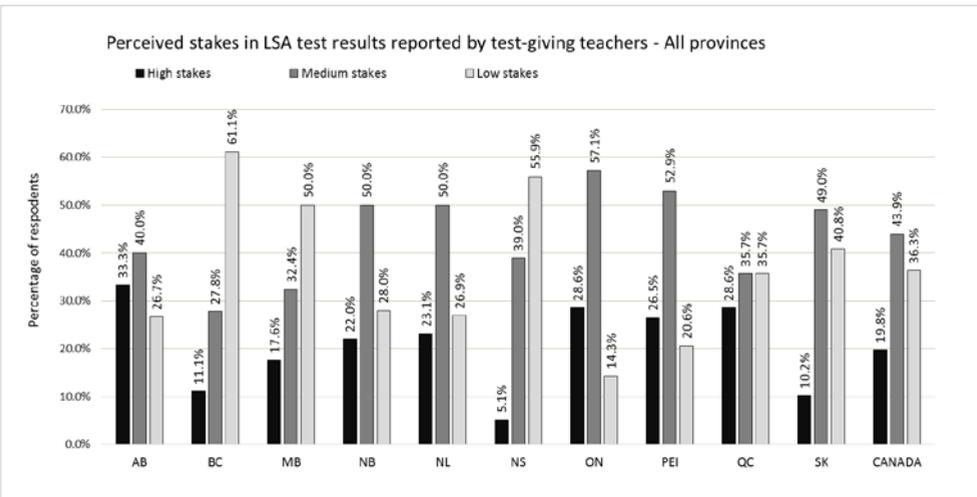


Figure 6.17: Comparing the stakes perceived by teachers who do not give LSAs and those who do.

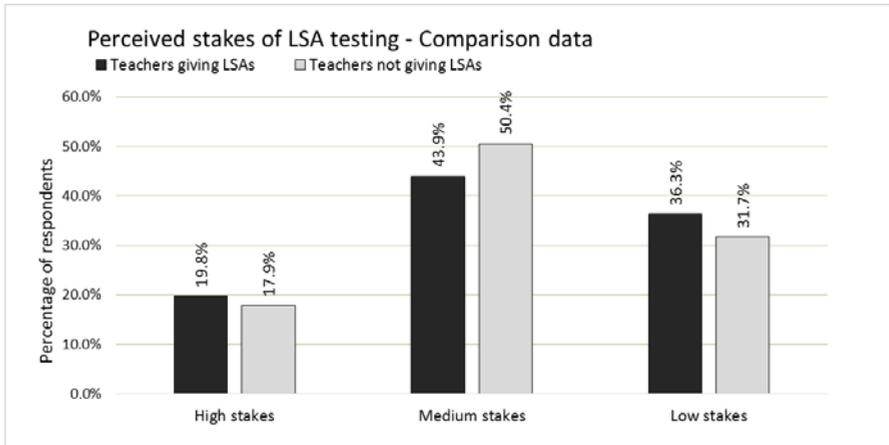


Figure 6.18: Residual analysis for incentives data and positive reactivity regressions

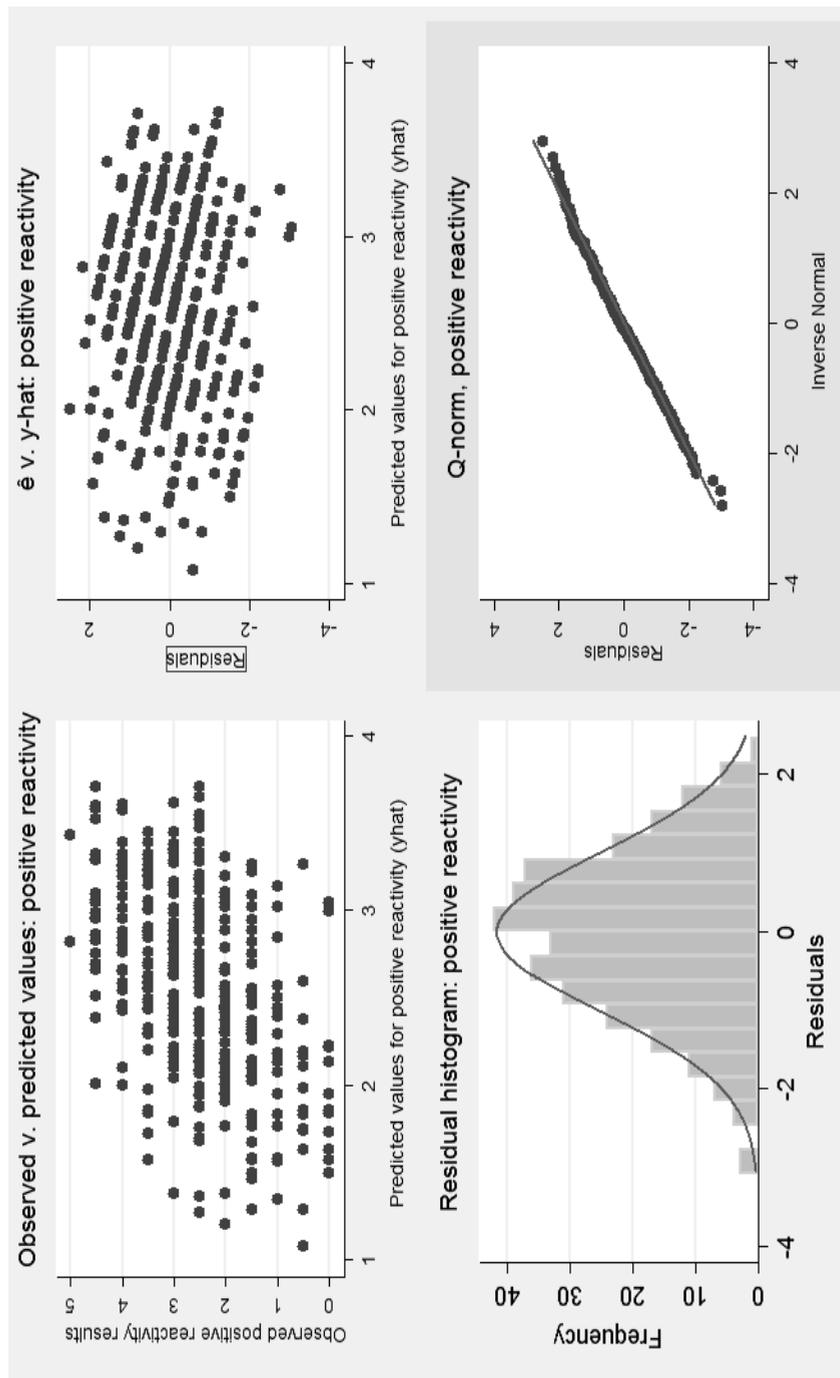
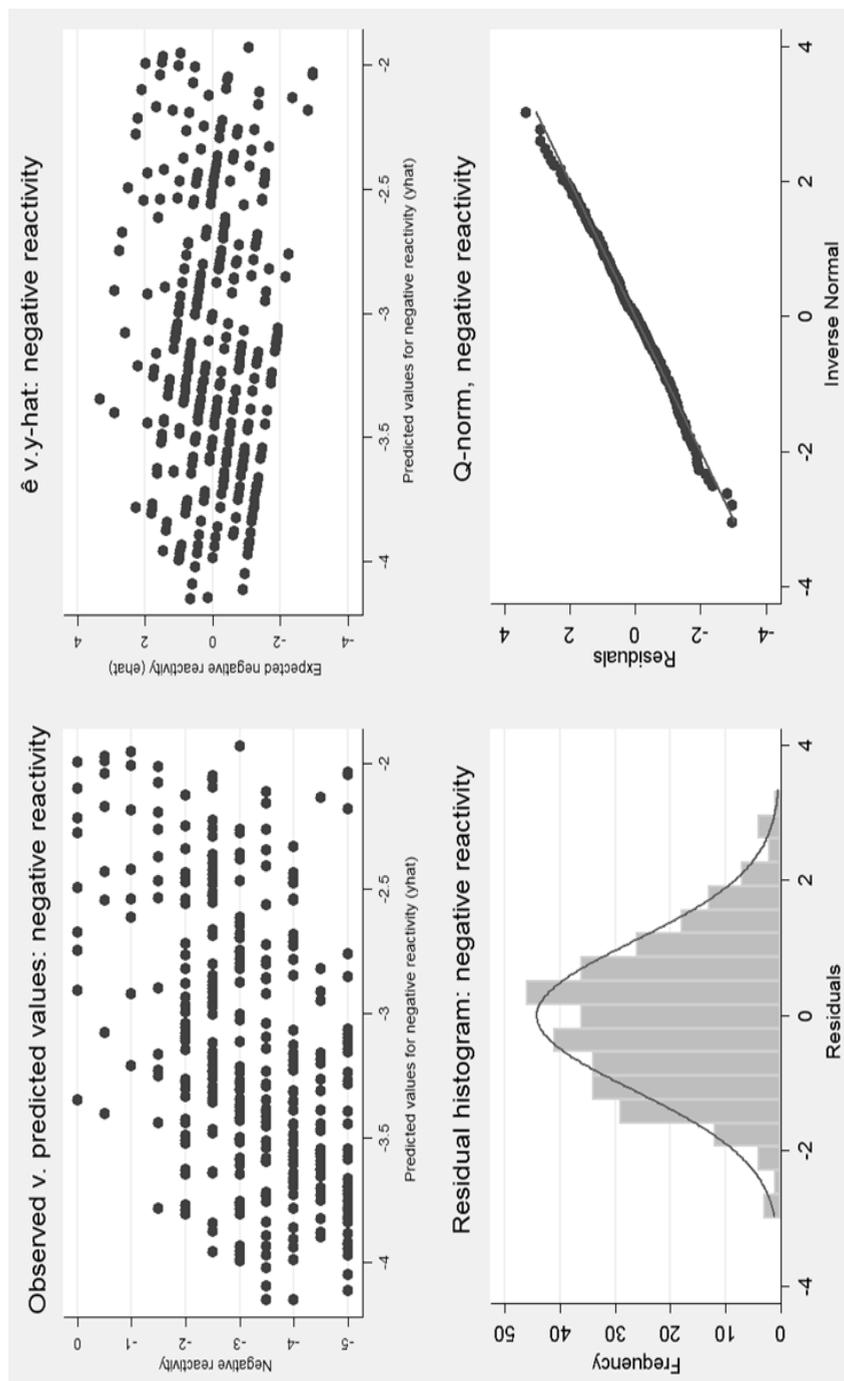


Figure 6.19: Residual analysis for incentives data negative reactivity regressions



Teacher background variables and data use

7.1 Introduction

If it were simple to nail down what individual factors made for an effective teacher, it likely would have been done long ago. This is not to say there have not been many attempts to define 'effective teaching' and to apply the lessons learned to both pre-service and in-service training. However, the current literature is forthright in saying that there is no clear consensus on what qualities or kinds of training make a teacher effective. The latest variant in this quest is value-added – an attempt to calculate how much a student's increase in test scores can be attributed to any one teacher. While debates rage about how to quantify good teaching, it will continue to happen, quantified or not, in our schools. The chapter that follows will examine background factors and their relation to the use of LSA data (reactivity) rather than their effect on results.

This chapter is laid out in the following way: (a) a literature review on teacher demographic and background variables and attempts to tie these same variables to student performance; (b) a presentation of the findings from the researcher's survey of teachers all across Canada; and (c) conclusions are drawn.

7.2 Literature Review

The literature regarding the characteristics of effective teachers has more ambiguity than that of any topic that has been examined in this dissertation. Just exactly which qualities make an effective educator would be great to understand but with our current level of understanding, effective teaching is too complex and thus defies quantifiable measurement and description. Knowing this, the author did not pursue this line of research to its fullest extent. Nor were the background questions in the survey intended to uncover some as-of-yet-undiscovered formula for what type of teacher is the most likely to make good use of quality data. This chapter was envisioned as a response to the criticisms that might rightfully be aimed at the study if these questions were not examined. Background factors are presented in order to try and dismiss the idea that there are any known specific qualities of teachers that will indicate they are more effective at using LSA data than any others.

It is also worth acknowledging that while the research cited in this review is related to 'effective teaching', this study is about teacher reactivity, a related but distinct perspective on the profession. Effective teachers might be those prone to use positive reactivity practices more commonly, but that is not what I have set out to study, nor will this chapter address reactivity as a metric of effectiveness.

The literature on effective teaching does not have many points of agreement, but a good place to start is with the literature on this topic, and that is the idea that teachers *do* matter. The differences between effective teachers and those that are less effective are of import to students and to schools. The Coleman Report (Coleman, 1967) was one of the first research papers to make this explicitly clear, even if it had been self-evident for a long time (Wenglinsky, 2002). Almost every paper that examines the quality of teachers makes reference to 'unobservable factors' that make it difficult to pin down which qualities exactly will make someone an effective teacher, and which should preclude them from the profession (Rockoff, 2003; Lytton & Pyryt, 1998; Borman & Kimball, 2004). Knowing what we do not know and cannot yet figure out, researchers, policymakers and school division human resources staff must use proxy measures to stand in these unknown quantities.⁹¹

Without any other metric, it is widely accepted that standardized tests are the only way to really measure effective teaching (Rivkin, Hanushek & Kain, 2005; Sanders & Horn, 1996). There are many points that have been brought up through this dissertation that would make this belief seem somewhat problematic. One primary consideration is that standardized tests are not assessments of teachers – they are assessments of their students, and second-hand metrics are not as accurate as first-hand ones (Koretz, 2008; Ravitch, 2010). Issues regarding test design, socio-economic factors for students, and inappropriate test preparation inevitably enter this discussion, and the belief that any one evaluative tool could account for these factors and still provide an accurate picture of what are admitted to be unobservable qualities seems almost quixotic. Yet since this is the metric that is used in research (and even the author will admit there is no other readily available metric at hand), it will permeate the literature review that follows since it is a foundational belief.

Returning to the idea that there must be some proxy measures available to replace what is not empirically defined, the most common stand-in used is years of teaching experience. It makes intuitive sense that if someone has been teaching for a longer period of time they would have insights and skills that younger, less experienced teachers would not have. Most studies into teaching experience have not supported this belief (Wenglinsky, 2002; Zigarelli, 1996). There is an acceptance that having some experience is beneficial, but it does not accumulate over an entire career, and it may even become a detriment if stagnation and aversion to change are the default positions of veteran staff members.

Next to experience, the most commonly cited stand-in qualities for effectiveness are educational credentials. Teachers with advanced degrees

⁹¹ This very idea, of using teacher characteristics to stand in for quality, is an issue raised by Jacob (2007).

(master's certifications) are widely assumed to be better teachers simply by virtue of these credentials. Pay scales in all Canadian provinces reflect this bias, as a master's degree will earn you a higher annual salary (depending in Québec on how many years of schooling in total you have been through) than a teacher with only a bachelor's degree. There are other educational credentials that can earn a teacher more pay as well, but these differ from province to province. The underlying assumption, that a teacher with more education is a more effective teacher, is not borne out in the research literature (Goldhaber & Anthony, 2005; Wenglinsky, 2002; Zigarelli, 1996). In some cases, studies have found that advanced degrees decrease effectiveness (Clotfelter Ladd & Vigdor, 2007).

When educational reform is discussed, class size is a common theme. It is not considered a stand-in for teacher quality but rather either a mitigating factor (to explain differences in assessment scores) or as a possible panacea to make wholesale improvements possible. The relative size of classes is a pertinent and important issue, but neither of these perspectives is much in line with what researchers have discovered. Class size, at least sizes that are practical in public schools (where 15 students is an almost impossibly low standard excluding research studies – as noted in Borman & Kimball, 2004), do not have much of an impact on standardized test scores, and as such, are assumed to not affect the quality of teaching (Wenglinsky, 2002; Willms, 2000; Hanushek, 2008).

A related issue which is often discussed in educational policy circles is what size of school is the most effective for students (Duncan & Noonan, 2007; Baker & Linn, 2002). A small school does not necessarily mean that classes are smaller, so it is not a co-variant factor to class size, but the variable does, according to current Canadian research, impact student achievement (Leithwood & Jantzi, 2007). This variable is therefore one that impacts the effectiveness of schools, but not necessarily teachers. School size seems to be linked more closely to the climate and culture of a school than to academics, even if there is some interplay between these factors themselves.

Another school-level factor that is thought to have an impact on achievement is autonomy (Santiago, 2002; Scafidi, Freeman & DeJarnett, 2001). The often-cited Finnish example is a good illustration of what this variable might do in terms of student achievement (Morgan, 2009). Charter schools in the United States, academies in the United Kingdom, 'earned autonomy' in the Netherlands, and private schools almost everywhere are put forward as examples of how giving schools more latitude regarding how they deliver the curricula may allow them to hit on more powerful and effective teaching techniques than those that are employed in the traditional public school system (Blok, Slegers & Karsten, 2008; and for another international perspective, see Wößmann, 2000). This variable appears in the literature, but it was not rated in this study's survey.

From the perspective of economists, the education production function is the most commonly accepted way to examine what happens in schools and in classrooms to determine what policies and practices are most effective. Inputs are such things as: (a) spending on resources; (b) more non-specific per-student spending; (c) upgrading facilities; (d) hiring more teachers; (e) teacher benefits such as pensions or health plans; or (f) paying teachers more based on qualifications or results. Aside from the last of these (incentives for better LSA results) the consensus seems to be that spending does not produce much in the way of improved LSA scores (Hanushek, 1997; Wenglinsky, 1997; Wenglinsky, 2000). The relative merits of incentive pay are yet debated and unresolved.

One last topic that is very much related is value added measures (VAM) of teacher quality. Going several steps more in depth than the production function equations do, VAM tries to account for external and internal variables to make LSA scores more informative about the quality of any given teacher. Debates rage about how useful these equations are (Koedel, 2009; Koretz, 2008), but the flaw in VAM from the author's standpoint is that whatever other factors are included in the complex equations used by VAM practitioners is that in the end they depend on LSA scores to provide a benchmark for what effective teaching should look like and LSA scores provide far less than academic consensus (Rockoff, 2003).

Figure 7.1: Summary of background factors literature

Topic	Author(s)	Summary Statement
Teacher quality	Borman & Kimball, 2004	This Nevada study set out to examine the link between teacher quality and equitable student outcome within any single classroom. Findings support a wide variation in teacher quality and that better teachers go some way to close achievement gaps.
	Clotfelter Ladd & Vigdor, 2007	Using North Carolina data to show credentials do matter: experience (50% of gains in the first few years, continues for an entire career), teacher licensing test scores, and regular licensure (regular certification identifies effective teachers). Degrees like a master's after 5 years lead to decreased student achievement.

	Coleman, 1967	Paper describing the history of educational quality as a changing concept in (mostly) American education. He is speaking mostly of de-segregation issues and the differences between schools for white students and schools for black students. This analysis asked important questions about what factors do make the most difference in student achievement including the quality of teaching which remains ill-defined.
	Goldhaber & Anthony, 2005	Results from the (US) National Board for Professional Teaching Standards certification test (National Board Certified Teachers qualification) are used to determine that these tests are an indicators of teacher quality but do not appear to promote professional growth.
	Hanushek, 1997	A meta-analysis method is used for the author to (again) examine the relationship between school resources and school achievement. He finds that teacher experience, class size, qualifications, salaries and other resource inputs do not provide clear or consistent gains in student achievement.
	Jacob, 2007	A paper on recruitment which delves into teacher quality issues. Jacob warns against using teacher characteristics as a proxy for teacher effectiveness since they are not the same, and have been shown to misalign in different ways.
	Morgan, 2009	Examines PISA history and its influence on Canada, and the Finnish model of high teacher qualifications and school-based pay-autonomy as model for what can improve scores.

	Koedel, 2009	This study examines student achievement in high school core subjects (Math, English, Social Studies and Science) in order to examine spillover effects (of learning from one subject to another) and also the strength of value-added measures. The author suggests caution using VAM including because they are not effective at calculating across-school variances.
	Koretz, 2008	Written as a response to the claims of value added measures advocates, Koretz argues that VAN is still very new, has lots of issues, and does erase the biases of test-based errors.
	Leithwood & Jantzi, 2007	A review of 59 post-1990 studies on school size indicating that smaller schools have comparable or better outcomes in academics, climate and culture, organization and cost efficiency, etc.
	Lytton & Pyryt, 1998	This 'effective schools' study in one district (Calgary) tries to explain between-school differences in achievement. The authors conclude that 50% is attributable to socio-economics, 10% is language (ESL), and 5% is teacher variables.
	Rivkin, Hanushek & Kain, 2005	Using data from Texas schools to determine the relative academic effects of reducing class sizes as compared to improving teacher quality. This does not specify what teacher quality is except to say it improves LSA scores and is not readily identified by experience, test scores, or other qualifications.
	Rockoff, 2003	Admitting that many aspects of teacher quality are 'unobservable' the author uses New Jersey schools' panel data to find that teacher fixed effects are strongly related to student achievement outcomes. By defaulting to what is observable (LSA scores) the author promotes the use of performance-based indicators of quality in hiring and determining compensation levels.

	Sanders & Horn, 1996	Using data from Tennessee, the authors argue that effective teachers promote much higher achievement gains from students and that the effects are additive and cumulative. Thus a student who gets strong teachers in succession sees vast gains over a student who is assigned poor teachers in succession. What it is that makes these teachers effective is not addressed.
	Schildkamp & Kuiper, 2010	A study examining how are data being used and what variables hinder or help. With data from the Netherlands characteristics of teachers who use data are mapped: data use skills; belief in the data; ownership/autonomy; and locus of control (belief that a teacher can make a difference).
	Wenglinsky, 1997	Looking at how schools spend their money, the author determines that some expenditures improve student outcomes (hiring more teachers), but some do not (higher pay for teachers, higher qualified teachers, paying school administration more).
	Wenglinsky, 2000	This paper uses 1996 NEAP data from students and teachers to question teacher input/output and their roles in increased achievement.
	Wenglinsky, 2002	The author effectively excludes many of the common default factors for teacher quality which might influence learning (i.e. experience, major, qualifications, and class size). He concludes that classroom practices do matter, specifically: a stress on higher order thinking; applying knowledge to unique problems; and more hands-on learning.

	Wößmann, 2000	TIMMS data are used to compare institutional differences across an international sample. The author finds that: central exams and standards improve outcomes; school autonomy for hiring and processes are positive, but for budgeting is not; and that teachers' union influence is also negative. Most comparable to this study were findings that scrutiny of assessment is a positive influence as is teacher autonomy related to instructional methods.
	Zigarelli, 1996	The author examines effective schools research by first placing studies into one of six 'constructs' and then compares these models for student achievement gains. Most effective are: achievement-based cultures; principal powers to hire and fire; and high teacher morale. One discredited construct is 'hiring quality teachers' based upon: education levels, in-service training; experience; verbal ability; prep time; and instructional strategies.

7.3 Preliminary hypothesis

H 7-1: Quality teaching is a difficult thing to quantify or measure and therefore it has not yet been determined to be a response to any given background variable or set of variables. For this reason the examined background factors related to the teaching population will not show significant effects on reactivity scores.

7.4 Results from surveys

This chapter will address several relevant background variables (BFVs):

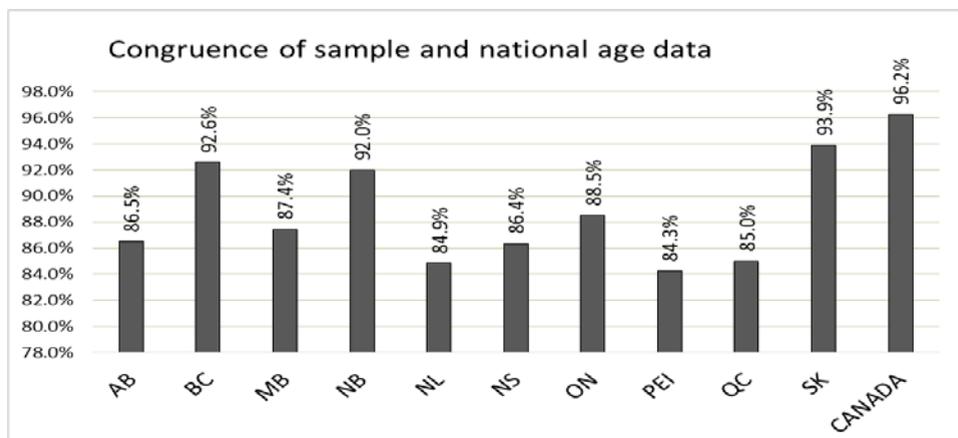
- BFV1 (background factor variable 1) – age
- BFV2 – sex
- BFV3 - grade taught
- BFV4 – experience
- BFV5 – setting (urban or rural)
- BFV6 - staff size
- BFV7 - class size
- and BFV8 – qualifications

Each of the background variables (aside from sex) was scored on an ordinal scale which was converted into cardinal values in order to perform the analyses. There were six possible choices for age, four choices for grade taught, six choices for experience, and so on. These groups and values are noted in **Annex 2**.

Only two of these variables could be compared with nationally available statistics, and these provided a means of comparing this sample with the larger teacher population. Analysing the alignment of the surveyed sample and the national population is done as a form of nonresponse analysis. Close alignment of these is an indication the sample is representative, at least with respect to the considered variables. Other data which are provincially available are nowhere near standard across different jurisdictions, and thus cannot be used to effectively compare provincial samples. These data are aggregated to the national level will be used in the regression analyses below to gauge their impact on reactivity effects.

The two background variables that can be compared to collected national data are age and sex of the teaching populations. The data from Statistics Canada⁹² has the ages of educators broken in groupings which were adopted by the researcher for the survey (**Figure 7.2**). To compare these national and sample data, the difference between the proportions of teachers at each age grouping was calculated and totalled to given a sum of differences. For example, the Canadian data show 1.9% of teachers younger than 25 years old, and the sample data showed 2.2% at this age level. The sample shows a 0.3% divergence from the national data.

Figure 7.2: Comparing survey respondents' ages to figures from Statistics Canada

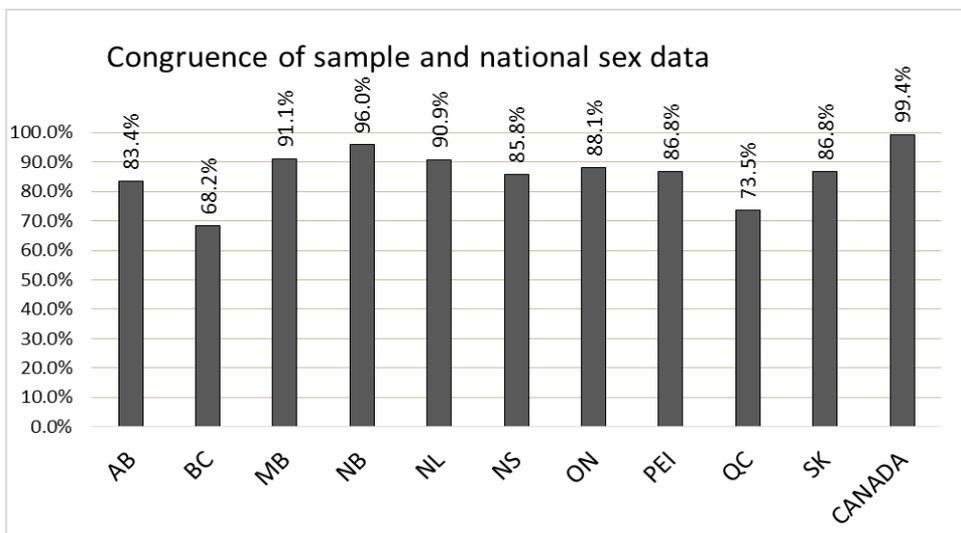


⁹² Age and sex data from the Statistics Canada. Retrieved Apr. 13, 2103 from http://publications.gc.ca/collections/collection_2007/statcan/81-582-X/81-582-XIE2007001.pdf

The absolute values of these divergences over all six age groupings were summed to calculate the total divergence, which subtracted from 100% provides the congruence figure (this same procedure was used to calculate the congruence of the setting data). In terms of the age data, the survey sample shows a 96.2% congruence in terms of the age with some variation in provincial samples. Note that higher variances in the provincial samples are a result of the smaller 'n' collected in each jurisdiction and that the use of more age group choices in the survey (there were 6 possible choices) leads to less congruence using this method. Provincial and national response rates are shown in **Table 2.1** in Chapter 2.

Figure 7.3 shows the congruence of sample and national sex data. The data collected from Statistics Canada breaks down the provincial and national teaching populations which can also be done with the sample research data. There is 99.4% congruence between the national population of teachers and the survey sample in terms of sex (the procedure for calculating this figure is discussed above). While some variation between provincial samples is evident, overall, the sample is strongly linked to the teacher population and indicates that based on the only two nationally known and measurable statistical evidence, the sample matches the population very well. According to Olsen 2006, such congruity shows that the sample is representative, at least with respect to these back ground variables.

Figure 7.3: Survey sample and national data on sex are compared



While reasonable samples were sought from all provinces in Canada, significant barriers were encountered that prevented the sample from being fully representative of the teachers in each province. The national data, aggregated as

they are, conform quite well to national norms. Some of the impediments that were encountered included:

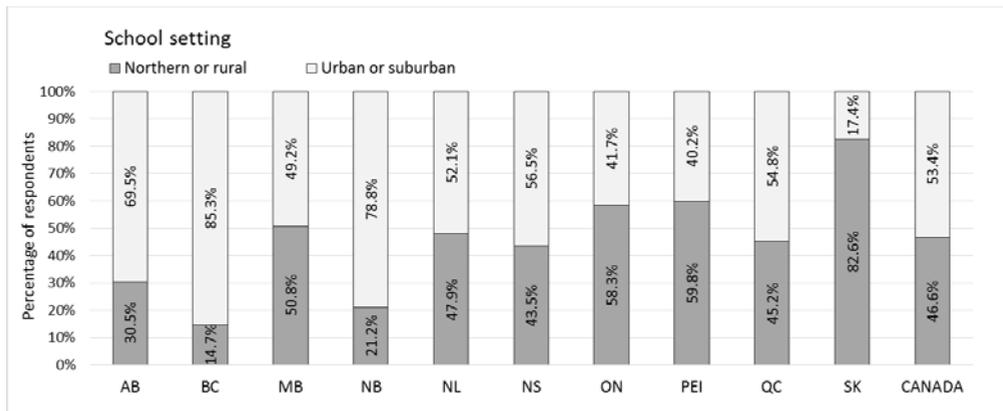
- the school divisions that denied access were mostly urban (they get many more such requests, especially if near a university), thus skewing the sample toward rural and remote respondents (the worst case of this was in Saskatchewan where both major urban centres denied requests, thus leaving no significantly large urban division left to sample)
- the number of declined requests to do research in Ontario (where 20 of 24 requests were denied and all to major centers)
- the large number of research requests that are passed on from universities to divisions and schools (it was not uncommon to hear of several surveys being sent to teachers during the same week)
- application barriers put up by some divisions (long online forms that are unique for each jurisdiction, requested application fees, requests for ten or more printed/mailed copies of applications and all supporting documents)
- processing times for applications varied from immediate to 3 months.
- teacher union job action in British Columbia that restricted access to schools in the spring and fall of 2014 (at this time teachers did not take emails from school administrators)
- weather prevented many schools in eastern Canada from taking part (there were upwards of 10 school days missed for blizzards, and time pressure was too intense to add anything else to the agenda)
- technical concerns including firewall blocking of the Survey Monkey page and the prohibition of using Survey Monkey in any capacity in the Nova Scotia public sector (including schools)
- the limited time that the researcher was able to commit to all the phone calls, emails and paperwork – to get a complete sample within one school year was barely accomplished

As a result, the data from the survey sample on school setting do not conform very well to some provincial data or, as a whole, to national figures. The main reason revolved around the researcher being denied access to urban/suburban school divisions and schools more frequently than was the case for rural/northern school divisions and schools (see **Figure 7.4**). Provinces that had the worst alignment of survey responses (by setting of schools) to provincial figures were Saskatchewan (where both major urban centres denied research requests), Ontario (where over 20 school boards denied access including all major cities), Québec (where no divisions from either Montréal or Québec City granted access) and New Brunswick (which has a unique over-representation of urban

respondents). There are, however, six provinces with quite good alignment regarding school setting.⁹³

The school setting data should be considered with some important caveats in mind. The data from Statistics Canada (StatsCan) gives figures for urban (81%) and rural (19%) population proportions were determined based upon provincial government standards for assigning 'city status.' Unfortunately, these standards vary from province to province and are not clear in the minds of many people, teachers included.⁹⁴ Not surprisingly, when respondents were asked to designate their school as urban or rural, teachers from the same school would often choose different options. Add to this fact that the urban/rural proportion standards used from StatsCan talk about where populations live and not where they attend school and we have another confounding factor. A rural student may be bussed into an urban school, for example. For these reasons, the data about location are not considered particularly reliable.

Figure 7.4: School setting data for respondents reflect the urban/suburban or rural/remote placement of teachers responding to the survey. Some reasons that this data does not align with the national figures are listed above.



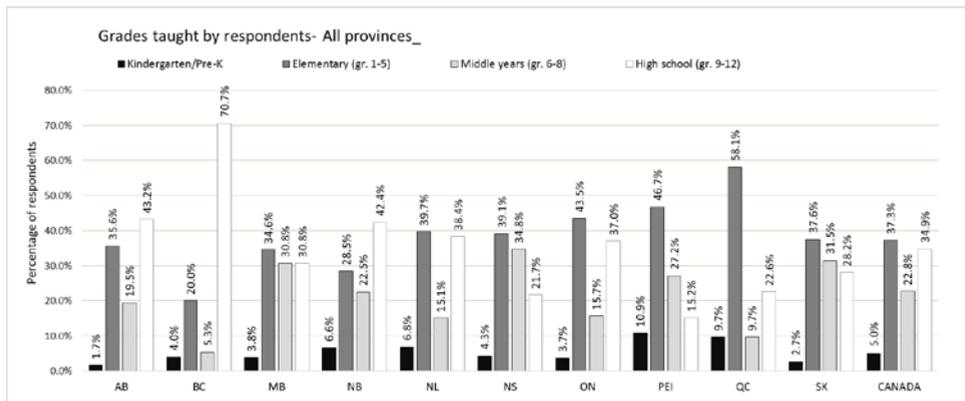
Looking next at another background factor, the provincial data do not have anywhere near equitable distribution of teachers across the grades (see **Figure 7.5**). That said, the national sample, like the sex and age data, evens out many of

⁹³ Setting data from Statistics Canada, retrieved Aug. 9, 2014 from <http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo62a-eng.htm>

⁹⁴ A city may apply for incorporation at a population threshold of 5 000 residents in Alberta, British Columbia and Saskatchewan, at 7 500 in Manitoba, and 10 000 in Ontario. In Québec the designation *ville* covers all communities regardless of population. The other provinces do not have specific limits for incorporation.

the rough patches to provide a fairly normal distribution pattern. The middle years numbers are low as a result of the 3-grade range it encompassed (the elementary grouping covers five grade levels, and the high school grouping covers four grade levels). Provincial response levels depended greatly upon the assent of administrators to distribute, and some principals were much more receptive to this research than others.

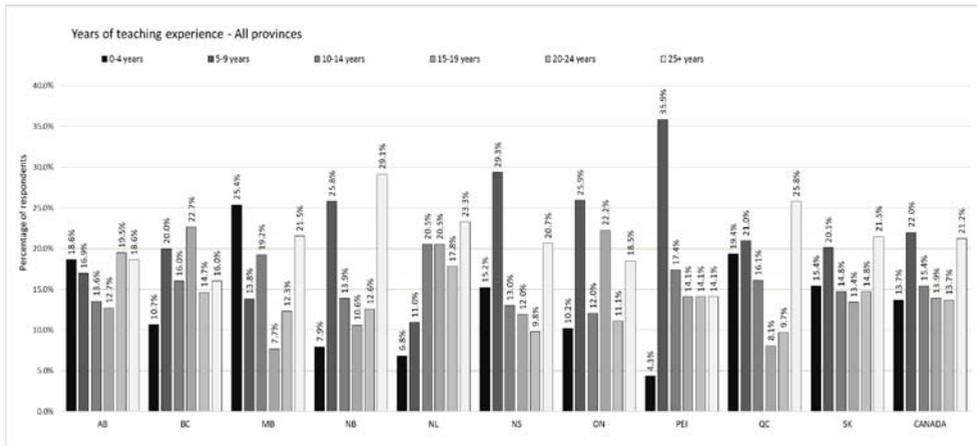
Figure 7.5: Grade levels taught by respondents



Next we see the self-reported data from respondents on their teaching experience (in **Figure 7.6**) which show variation between provinces, but not in such a way that any grouping (teachers in the early stages, middle stages, or toward the end of their careers) dominated the dataset. There are low numbers of very young teachers (0-4 years) in New Brunswick (7.9%), Newfoundland and Labrador (6.8%), and especially in Prince Edward Island (at 4.3%, but with a large cohort of teachers in the second experience tier from 5-9 years). Government policies for recruiting and retirement as well as university education program graduations rates each play a role in determining the number of teachers at these levels.

Where a large number of young and very young teachers are evident, there is commonly offsetting numbers of older and end-of-career teachers (20-24 years and 25 or more years, respectively). This is seen in Alberta, Manitoba, New Brunswick, Nova Scotia, Ontario, Québec and Saskatchewan. It is proportionally the second highest respondent group in the national sample (21.2%) and the largest group in New Brunswick, Québec, Newfoundland and Labrador, and Saskatchewan. This may be indicative of a lot of provincial hiring in order to meet student-teacher-ratio standards, or hiring to make up for retiring baby boomers. It may also be partly a result of teachers working past their 30th year, which is commonly the tap-out point for retirement with full pension benefits.

Figure 7.6: Years of teaching experience reported by survey respondents



Where it is permitted, teachers can 'double-dip' by receiving their pensions and continuing to work teaching contracts (that are in some cases restricted to less than a full year to qualify for pension benefits). A distribution analysis for this group would not be informative since it is not, and should not be, normally distributed.

Examining school size (see **Figure 7.7**), the provinces that have larger rural populations also tend to have smaller schools (school and divisional amalgamation has occurred since the 1980s in various provinces, but the populations of rural areas still demand some school facilities be located close to home) and smaller staff sizes (note that Manitoba, Newfoundland and Labrador, Nova Scotia, Prince Edward Island and Saskatchewan all have higher numbers of small schools – these are five of the six most rural provinces by ratio of the population). It is true in this case as well that the national sample tends to even out the provincial variations and presents a fairly uniform set of proportions for school size.

It is perhaps a result of the small steps (of 10 teachers) from one level to the next that so many jurisdictions show larger tails in these distributions. Nova Scotia, Saskatchewan and Prince Edward Island, in particular, show this effect, which is thought to be the result of small rural schools co-existing with large consolidated schools which have large staffs to accommodate the needs of the students travelling in to school from communities without schools. This is especially true at the high school level where students are considered old enough to spend time riding a bus (or driving) to and from school, and when having specific course options available, university is for many of these students just around the corner, is a key consideration.

Figure 7.7: School size as reported by survey respondents

Class size (see **Figure 7.8**) is a common theme in school improvement literature, but here the national distribution is fairly normal with very large classes (30 or more students) most evident in Alberta (27.1%) and Nova Scotia (12.0%), and very small classes (1-10 students) evident in only very small proportions in seven provinces, nationally rated at 2.2%. Classes of 21-25 students were the most commonly indicated (40.4%). The distribution (see **Table 7.9**) is analysed below. The distribution is skewed right (-0.4126) and somewhat stretched (kurtosis: 3.3067).

Figure 7.8: Average class sizes as reported by survey respondents

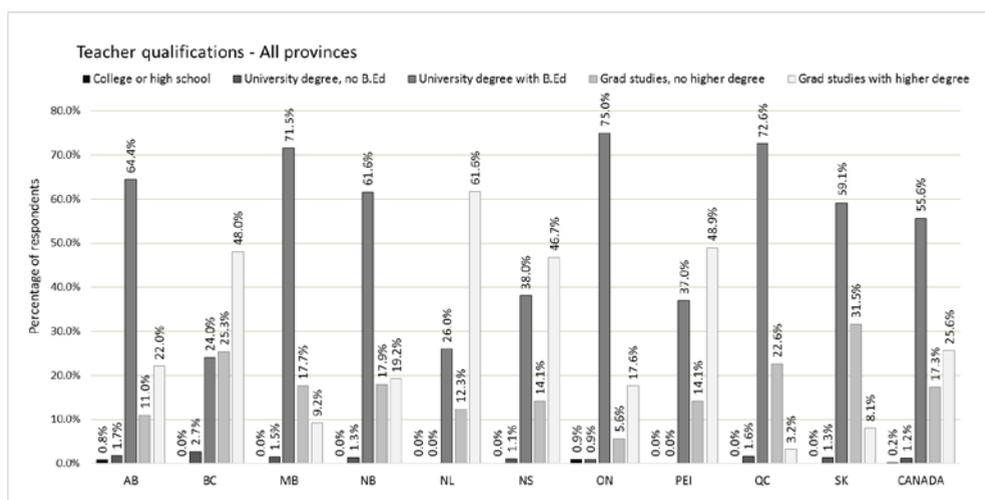
Table 7.9: Distribution analysis for the national data in figure 7.8

Class size		Observations: 1048
Mean: 3.9924	Standard deviation: 1.0443	Variance: 1.091
Range: 1 to 6	Skewness: -0.4126	Kurtosis: 3.307

The variable to mean ratio (D value) is quite low at 0.2732, meaning that the distribution is significantly under-distributed. This is likely partly a result of the small range of possible values (six groupings).

Contract terms for teachers vary across the provinces of Canada, and this does have a significant effect on the qualifications that are most common in each jurisdiction (see **Figure 7.10**). Some provinces do pay for higher qualifications, and the extra salary seems to be a sufficient inducement to have many highly qualified staff (with graduate degrees) in Newfoundland and Labrador (61.6%), Prince Edward Island (48.9%), British Columbia (48.0%), and Nova Scotia (46.7%).

Figure 7.10: Qualifications data from survey respondents



In all other jurisdictions, staff with Bachelor of Education qualifications is the largest group, and this is true nationally as well (55.6%). The number of respondents without no university credentials or having less than a university B.Ed. degree was very small proportionally (together only 1.4%).

7.5 Correlation analysis – background factors

As in previous chapters, the relationships between the background factors are examined here using Spearman's rank order correlation tests (see **Table 7.11**). These correlation tests are helpful in determining which factors are most closely aligned with each other before the regression analysis is done.

Of the several interesting correlations evident here, some make sense intuitively, some take pondering to decipher.

Age: By far and away the strongest correlation is between age and experience (a value of 0.783, where the variables had larger values attached to increased ages and years of experience). Male teachers appear to be younger in this sample, and older teachers also tend to be better qualified (higher values were assigned for more qualifications).

Sex: While the sex variable (female teachers were coded as 2, males as 1) is significantly correlated with several other factors, the strongest relation is to grade taught (higher values were assigned to higher grades) since elementary teachers are predominantly female. All the sex correlations are negative showing that the male respondents also reported less experience, smaller staff and classes, and fewer qualifications.

Grade taught: The grade taught factor is strongly correlated with staff and class sizes (higher values were given for larger staffs and classes), and also with higher qualifications. Higher grade teachers also tend to be more commonly located in urban/suburban schools.

Experience: Beyond what has already been noted, this factor is also positively correlated with more qualifications.

School setting: Urban/suburban schools tend to have larger classes and staffs than rural/northern schools (urban/suburban were coded as 1, rural/northern as 2).

Staff size: Very apparent in higher grades, large staffs (higher values were assigned to larger staffs) seem also to correlate with urban/suburban schools, with large class sizes and be more widely reported by men.

Class size: This is a mirror image of the staff size data – all the same correlations exist in virtually the same order (higher values were assigned to larger class sizes).

Qualifications: Finally, higher qualifications are correlated with advanced age, more experience, and teaching higher grades (higher values were assigned to more highly qualified teachers). Qualifications are negatively correlated with sex, so are less common for men.

Table 7.11: Spearman's rank order correlation test done with background factor variables

Correlation matrix - background factor variables

1. Age	1.0000							
2. Sex	-0.085**	1.000						
3. Grade taught	-0.028	-0.282**	1.000					
4. Experience	0.783**	-0.094**	0.034	1.000				
5. School setting	-0.004	0.040	-0.119**	0.013	1.000			
6. Staff size	-0.037	-0.129**	0.607*	0.005	-0.435**	1.000		
7. Class size	0.018	-0.136**	0.464**	0.028	-0.402**	0.483**	1.000	
8. Qualifications	0.152**	-0.091**	0.098**	0.188**	-0.011	0.016	0.042	1.000

* p<0.05; ** p<0.01

The correlation matrix for background factors shows a high number of significant relationships. As independent variables, the relationships are strong and border on being concerns for collinearity, the highest two values being 0.783 and 0.607.

7.6 OLS regressions – background factors

7.6.1 Regression analysis

The background factor variables collected in the survey are those commonly examined in the literature as well as those available from Statistics Canada. Dummy variables were added to examine provincial variation, but note that PEI is the control province (it does not have dummy added) for total reactivity, BC for negative reactivity, and MB for positive reactivity.

There are two variables that have significant impacts on positive reactivity effects (see **Table 7.12**). The **age** results suggest that the older the teacher, the less likely it is that they will use positive reactivity strategies. This result shows significance at the $p < 0.05$ confidence level. Another significant correlation is that the larger **class size** is, the more likely a teacher is to use positive reactivity practices. This finding is also significant at the $p < 0.05$ confidence level. Prior to the addition of provincial dummies, these variables account for just 4% of the variance in survey response scores.

Table 7.12: Background factor variables and positive reactivity

Positive reactivity

Age	-0.238 (2.50)*	-0.198 (2.12)*
Sex	-0.082 (0.60)	-0.094 (0.69)
Grade taught	0.008 (0.09)	0.069 (0.74)
Experience	0.114 (1.94)	0.090 (1.57)
School setting	0.169 (1.14)	0.251 (1.66)
Staff size	-0.027 (0.46)	-0.015 (0.26)
Class size	0.171 (2.45)*	0.169 (2.34)*
Qualifications	0.004 (0.07)	0.021 (0.29)
(provincial dummies) AB		0.496 (1.87)
BC		-0.415 (1.46)
NB		0.616 (2.45)*
NL		0.330 (1.11)
NS		0.081 (0.30)
ON		0.485 (1.73)
PEI		-0.105 (0.43)
QC		0.573 (2.00)*
SK		-0.209 (0.69)
Constant	2.184 (3.44)**	1.560 (2.36)*
R²	0.04	0.12
N	368	368

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

Despite the very small R^2 value and thus the small effect leverage here, both of these are somewhat surprising findings. The **age** variable was really only collected as a means of comparing the survey sample to available population data. It was not expected to show significance especially in light of the fact that a highly correlated but distinct variable, years of experience, does not show significance. High levels of data use by younger teachers may also be a factor in this result, perhaps a function of their level of comfort with data, or the fact that most testing systems are relatively new developments, and thus were initiated during formative teaching years. The use of data by teachers with larger classes was also unexpected, but may indicate that teachers with comparatively large (but not extremely large) classes are somehow more adept at using data or have more impetus to use the data at hand, including LSA results. It could also be the case that teachers with smaller classes rely less on external assessment measures. This may be a result of small school lacking the internal expertise or organizational strength to deal well with these data.

Both New Brunswick and Québec show weak positive correlations with the control group (Manitoba does not have a dummy, and is the stand-in control), and thus show somewhat more positive reactivity based on background factors. Including the provincial dummy variables does increase the R^2 value of the regression to 12%.

Looking at the **negative reactivity** results (Table 7.13), we see that there is only one significant relationship. It appears here that the higher the **grade taught**, the more likely a teacher is to use negative reactivity strategies. This correlation is significant at the $p < 0.01$ confidence level, stronger than both of the positive reactivity correlations and responsible for 6% of the variance in scores prior to the addition of provincial dummy variables.

This finding indicates that high stakes exit exams from high school are likely important factors in the use and application of test-specific strategies (negative reactivity) that should result in higher scores. It follows that, at the national level, teachers of elementary and middle years grades are less inclined to use these strategies than high school teachers. These findings point to the fact that high stakes exit exams are a significant contributor to the adoption of negative reactivity strategies while they do much less to promote positive reactivity practices (support in the literature can be found in Abrams, Pedulla & Madaus, 2003; Koretz & Hamilton, 2003; Yeh, 2005).

There are two significant provincial correlations the first being strong and positive showing that Nova Scotia teachers report less negative reactivity than is true in the control group. The second correlation is weak and negative indicating that PEI teachers reported using somewhat more negative reactivity strategies than the control group (BC in this regression). Inclusion of the dummy variables does increase the value of the R^2 figure to a very healthy 21%.

Table 7.13: Negative reactivity effects examined in light of background factor variables. **It is important to note that since negative reactivity is enumerated in negative integers, a negative coefficient means more negative reactivity effects, not less.**

Negative reactivity

Age	0.062 (0.61)	0.088 (0.93)
Sex	-0.111 (0.76)	-0.133 (0.97)
Grade taught	-0.249 (2.55)*	-0.297 (3.11)**
Experience	0.000 (0.00)	0.028 (0.47)
School setting	0.068 (0.43)	0.027 (0.17)
Staff size	-0.068 (1.12)	0.000 (0.01)
Class size	0.016 (0.22)	-0.085 (1.15)
Qualifications	0.090 (1.26)	-0.004 (0.06)
(provincial dummies) AB		-0.195 (0.73)
MB		0.523 (1.81)
NB		-0.233 (0.89)
NL		-0.404 (1.35)
NS		1.046 (3.84)**
ON		0.096 (0.34)
PEI		-0.571 (2.13)*
QC		-0.528 (1.74)
SK		0.302 (0.91)
Constant	-2.701 (4.03)**	-2.053 (3.01)**
R²	0.06	0.21
N	369	369

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

The effects of more negative reactivity in higher grades explains some 6% of the variance in scores, but adding the provincial variations increases that value to 21% and verifies the finding that high stakes exit exams have a role to play in this significant figure (you will note the negative coefficient showing more negative reactivity in AB, NL, QC, and BC as the control, is effectively '0').

For **total reactivity (Table 7.14)**, we see the re-emergence of the same three factors analyzed in previous sections, and the correlations are all significant only at the $p < 0.05$ confidence level. It is clear (as also pointed out regarding positive reactivity) that the higher a teacher's **age**, the less likely they are to employ any reactivity strategies in their practices related to LSAs. This might be understandable in terms of what older teachers are willing to take on late in their careers, but it does not explain the fact that more years of experience remains non-significant though closely correlated to the age variable.

Class size appears again in this regression as a weak and positive relationship with total reactivity. As previously discussed, there are several factors that may well align with class size to create this positive correlation (larger schools and teaching staffs, larger classes being more common in higher grades, larger classes being more evident in urban/suburban settings, etc.). It is hard to imagine that simply having larger classes leads to more reactivity without the influence of at least some of these correlated factors.

The **grade taught** factor, highlighted in terms of its effect on negative reactivity, remains significant in the examination of total reactivity effects and is the strongest correlation of these three. It is only significant after provincial dummies have been included which may be a bump in significance based upon provincial variations and exit exams (as above).

Yet the effects of these factors appears limited and are responsible for only 5% of the variance in total reactivity scores prior to the inclusion of provincial dummies. When the dummies are included in the analysis, the R^2 value jumps significantly to 17% and four provinces also show significant results, diverging from the PEI control group. Nova Scotia has a strong negative correlation indicating less total reactivity, and Saskatchewan, Manitoba and British Columbia all have weak negative correlations which show a similar divergence from the control group. The obvious factor to point to in all of these cases is that the total reactivity averages for these four provinces are the lowest four nationally.

The independent variables in this chapter have only weak correlations with reactivity effects of all types. It is also the case that just three of the background variables out of eight show any significant effects at all. Several provinces show weak correlations with the control group, all indicating less total reactivity, but only Nova Scotia's correlations were strongly significant and seem rooted in the low levels of both negative and total reactivity evident in this province.

Table 7.14: Total reactivity and background factors

Total reactivity

Age	-0.321 (2.01)*	-0.304 (1.98)*
Sex	0.050 (0.22)	0.055 (0.25)
Grade taught	0.259 (1.69)	0.376 (2.45)*
Experience	0.126 (1.28)	0.065 (0.68)
School setting	0.042 (0.17)	0.171 (0.68)
Staff size	0.022 (0.23)	-0.036 (0.37)
Class size	0.159 (1.35)	0.260 (2.18)*
Qualifications	-0.073 (0.63)	0.051 (0.42)
(provincial dummies) AB		0.195 (0.49)
BC		-0.897 (2.06)*
MB		-0.975 (2.40)*
NB		0.454 (1.23)
NL		0.217 (0.48)
NS		-1.543 (3.75)**
ON		-0.113 (0.27)
QC		0.621 (1.45)
SK		-0.988 (2.14)*
Constant	4.964 (4.64)**	4.110 (3.81)**
R²	0.05	0.17
N	360	360

Tables show regression coefficients; *t*-values of the regression coefficients are bracketed; * $p < 0.05$; ** $p < 0.01$

7.6.2 Residual analysis – background factors

The residuals from these regressions were examined using four different econometric graphing techniques and the results from these analyses were fairly uniform across all three regressions. The results for the **positive reactivity** residual examinations are found in **Figure 7.15** (and at the end of the chapter see **Figures 7.20** and **7.21** for negative and total reactivity analyses). The 'observed v. predicted values' chart has a weak linear trend as it should since these variables show little effect on reactivity. The bands seen in the graph indicate different reported levels of reactivity (from 0 to 5 by multiples of 0.5). The ' \hat{e} v. \hat{y} ' chart has these same bands, but no clustering or significant outliers. The residual histogram has a relatively normal distribution and is thus more likely to meet the assumption of normally and independently distributed residuals required for hypothesis testing in OLS analysis. Finally, the QQ plot indicates small deviations from the normal distribution at the tails quite clearly. This quantile analysis checks the tails of the distributions in particular and some deviation is expected in these data as a result of discrete variables being used. In all, these analyses bear out the rigour of the regression model and help to confirm the findings.

7.7 An inquiry into subject area qualifications

Respondent teachers were asked about their university qualifications, and also about the subject of the LSAs written by their students. The four figures that follow illustrate that the number of teachers with qualifications in the subject area of the given LSA is quite variable. These figures (**7.16 – 7.20**) indicate first how many teachers responded that they give these subject-specific tests in their classes. For those who responded positively, their university major (major course of study) and minor (secondary course of study) were plotted to see if they align with tests they administer in their classes. In some cases it was necessary to clarify the actual number of teachers as a result of some science majors, for example, also having a science minor.

We see first for English LSAs, the number of teachers who report giving tests ranges from 2 in kindergarten, to 125 in elementary grades, 82 in middle years grades, and 87 in high school (see **Figure 7.16**). The number of teachers with English qualifications is quite low in elementary and middle years classes which also are tested. Including both majors and minors, 42.4% of tested elementary classes and 42.7% of tested middle years classes had English-qualified teachers. The proportion was much higher for high schools, but even here only 65.5% (57 respondents out of the 87 tests reported) of teachers giving English LSAs have any university English credentials.

Figure 7.15: Residual analysis for background factors and positive reactivity regressions

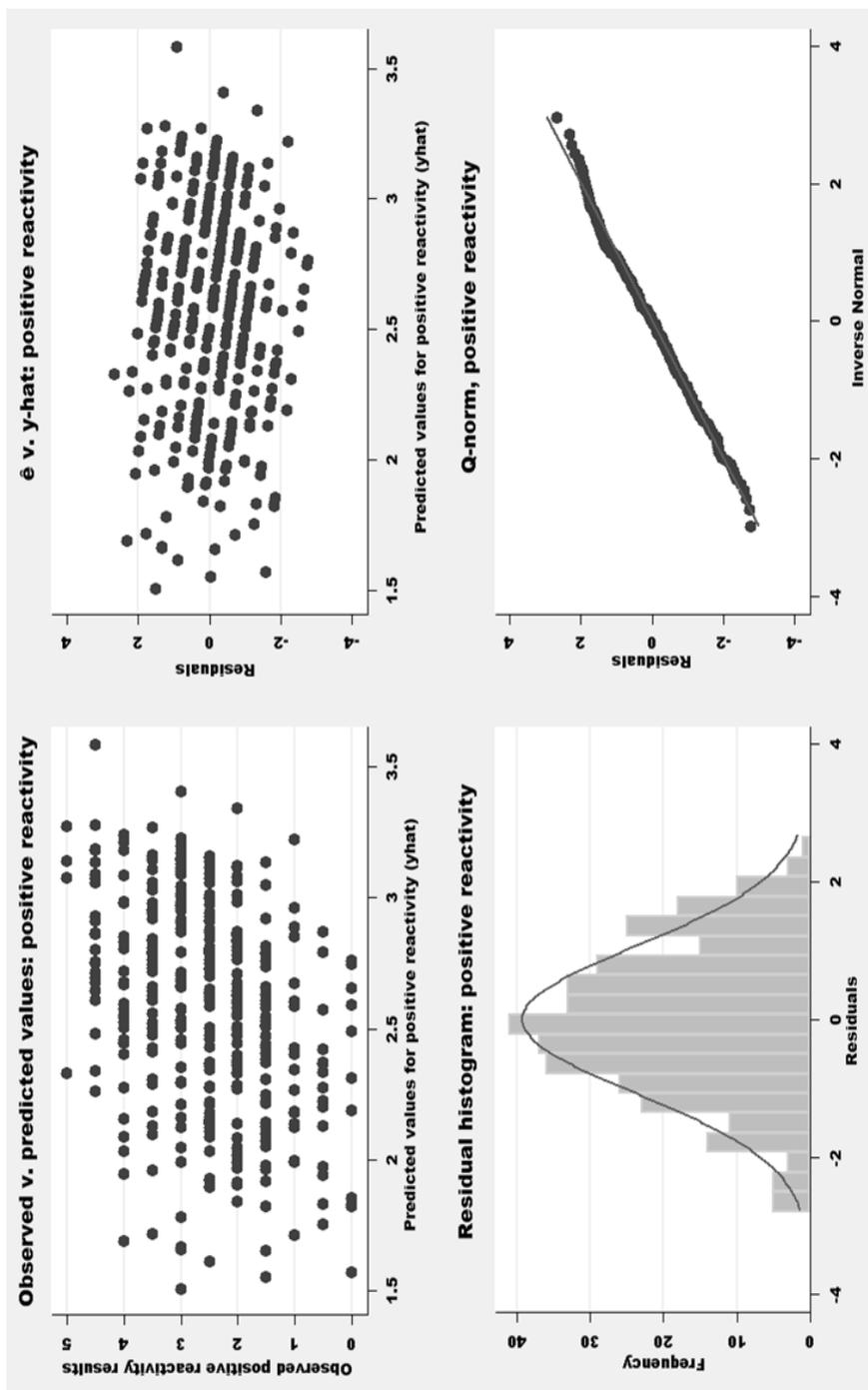
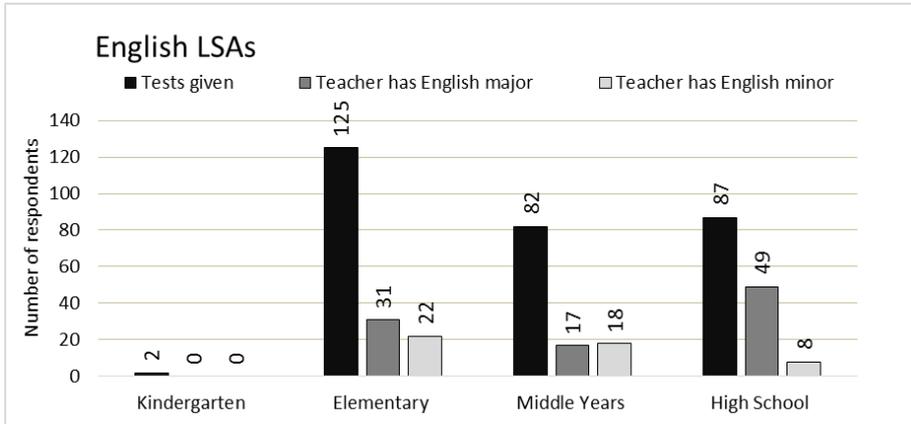
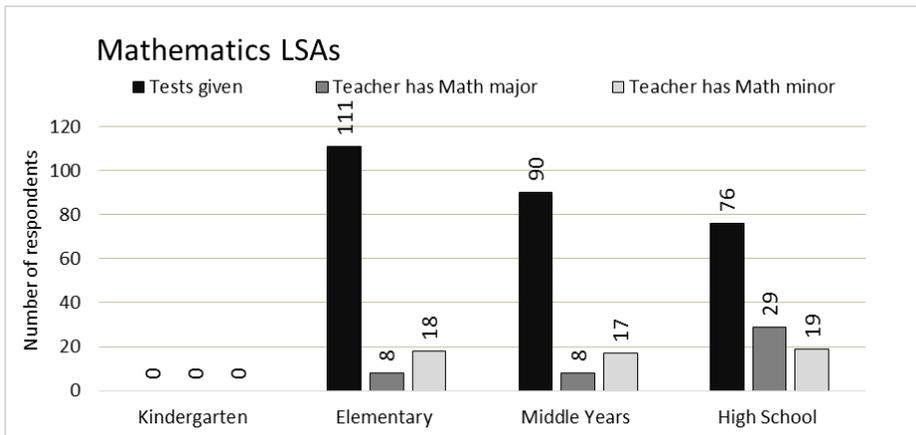


Figure 7.16: Teachers who give English LSAs and their qualifications



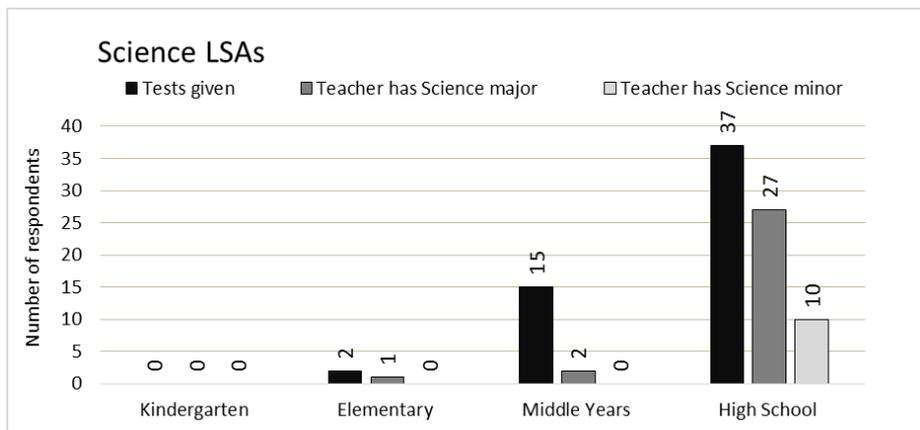
Math survey results also show very few teachers qualified in this subject area for younger students (see **Figure 7.17**) in classes that give LSAs. Only 23.4% of elementary teachers who give math LSAs have university level math credentials. For middle years the number is 27.8%. It is only in high school that more than half these teachers have the qualifications in mathematics. It should be said that 63.2% of high school math teachers (a very similar proportion of qualified teachers to English) still leaves more than one in three classrooms at a disadvantage in terms of subject-specific content knowledge (see Smith, Desimone & Koji, 2005 for an American study that examines teacher's math content knowledge and teaching practices).

Figure 7.17: Teachers who give mathematics LSAs and their qualifications



Science classrooms in high schools seem best equipped to handle the content and the science assessments at this level (see **Figure 7.18**). After removing teachers with their major and their minor in science (seven individuals who would otherwise be counted twice), 81% of teachers giving science LSAs reported having a science major or minor.⁹⁵ There were significantly fewer science assessments in earlier grades, but only 13.3% of middle years teachers giving science LSAs reported university credentials in Science. The elementary sample was at 50% but from a very small sample (two science LSAs were administered in these grades).

Figure 7.18: Teachers who give science LSAs and their qualifications

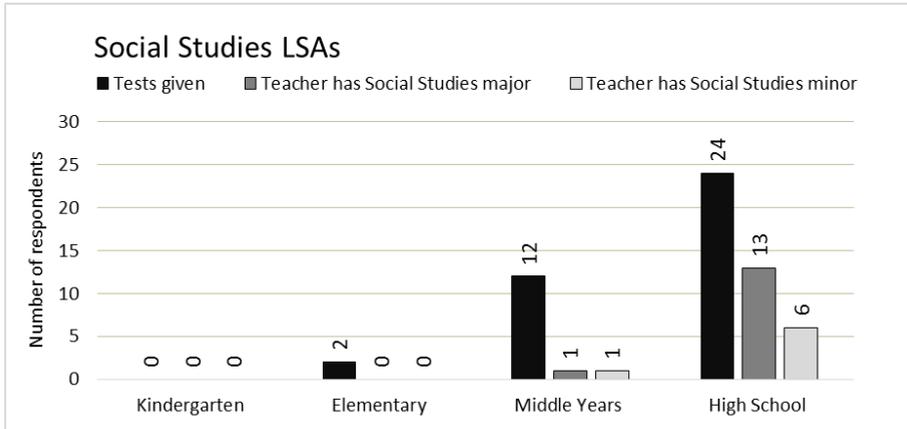


The final 'core' subject in the analysis of credentialing is social studies. This group of related studies (history, geography, etc.) appeared to have the fewest number of respondents indicating that their classes wrote provincial tests (**Figure 7.19**). Just 38 teachers reported giving LSAs, but it appears in these data as well that the number of university-qualified social studies teachers is quite low. The elementary sample is too small for solid conclusions, but only 16.7% of middle years teachers were teaching in their area of training as compared to 79.2% of high school social studies teachers.

Having determined the relative level of qualifications in these subjects when LSAs are administered, it is possible to turn now to reactivity effects. Analyses were done using one explanatory variable and intercept in a regression for one dependent variable. Whether in terms of total reactivity effects, positive effects or negative effects, the results are the same across all four core subject areas examined by this survey.

⁹⁵ Science was the only core subject where any respondents giving LSAs in that subject reported having both a major and minor in the discipline.

Figure 7.19: Teachers who give social studies LSAs and their qualifications



There are no significant correlations for any reactivity effects for subject-area-qualified teachers as compared to non-subject-area-qualified teachers. The *p*-values of the *t*-statistics are always larger than the 5% significance level and most adjusted *R*² values were below 1%. While subject-area credentials may well have an impact on LSA scores, they do not appear to have any significant correlation with reactivity effects.

7.8 Conclusions

The hypothesis that background factors are not significant actors in the reactivity dynamic has been shown to be less solid than had been presumed. There are very few background factors that rate even as blips on the radar regarding reactivity effects but three do register as significant factors: age, grade taught, and class size. Age and class size are significant at the $p < 0.05$ confidence level for positive reactivity. Grade taught is significant at the $p < 0.01$ confidence level for negative reactivity. Age, grade taught and class size all reappear as significant looking at total reactivity, and all are at the $p < 0.05$ confidence level. The relative strength of these factors is also important. The *R*² values for all background factors after including provincial dummies were: positive reactivity, 12%; negative reactivity, 21%; and total reactivity, 17%. While some of this is surely the result of provincial variation, the inclusion of provincial dummies indicated that none of these variations save Nova Scotia's were strongly significant.

This being a single study of reactivity effects (and the only one known by the author conducted in Canada), it is unlikely that these relatively low figures for both significance and strength will serve to convince the research community that

either age or class size are key factors in the potential use of LSA data and educational reactivity to large-scale assessment.

The finding that high stakes exit examinations seem to be more likely to promote the use of negative reactivity practices, on the other hand, is more commonly found in the literature (Darling-Hammond, 2003; Finnigan & Gross, 2007; Firestone, Mayrowetz & Fairman, 1998) and the regression result is somewhat more telling.

Finally, there appears to be no correlation between having subject area credentials and reactivity effects. There are certainly many teachers who give LSAs in subjects for which they lack university level training (which is another issue in itself), but this does not seem to have an impact on whether or how the results are used to improve instruction.

7.9 Final words

To end this chapter which examines what 'teacher qualities' might lead to the best results, I will leave the last words to some interview respondents whose insights have helped make sense of these data. Their knowledge and experience of local conditions in all of Canada's ten provinces was invaluable to the author and deserves all of the credit for moving this work from somewhat cold econometric analyses to the human scale.

Testing and teaching are not synonymous:

We, as classroom teachers, like to be able to look at an exam and say, 'my kids did alright on this thing.' That is reassuring. But other than that, I mean. . . It is an after the fact ranking and sorting that's happening of kids which is not . . . despite some people's expectations, is not what I think high school is for. So you know, go ahead and keep the diploma exam to rank and sort kids for whatever. . . But I am a classroom teacher. It is my job to teach. It is my job to get them to learn. It is not my job to rank and sort. So the purpose of that diploma exam and my purpose are very different. - **AB, High school Math teacher, female**

Teachers know their students' strengths:

I think administrators know how their schools are doing, how their students are doing, and they could easily ask the administrators. I don't think we have to put the kids through this testing to find out how they are doing. . . It is a waste of money, it is a waste of teaching time when you think of all the hours the kids are writing these assessments.

- **BC, Elementary homeroom teacher, female**

Ongoing professional growth and collaboration improve teaching:

No. I'll be blatantly honest, no. I think teaching improves through good collaboration, good, strong administration, the willingness on teachers part, collaboration with parents . . . and good professional development. I don't think any test is going to, you know, teach teachers how to teach differently or better. -

MB, Elementary school principal, female

These data can be informative to teachers and students:

I would personally like to see [more provincial testing]. I know as a school here, we collect all of the mathematic, umm, 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

What kinds of feedback really helps teachers:

How the heck do I know where I'm falling down on the job if I can't get some feedback? And that [provincial testing] is really one of my key features for feedback. So how do I improve from year to year if I don't know how effective what I delivered was? You know I can ask the kids 'How was that?' . . . 'Well that was fine, sir' - they just want to get off to lunch.

-

NL, High school Science teacher, male

Data-informed decisions can help schools and students:

The other thing I think, for us in the . . . school board that has I think been a positive outcome from the assessment process is that we have actually looked at the results and used them for instructional purposes and for developing interventions. . . So there was direct correlation between those results and action that has been put in place. -

NS, Division staff, female

Some barriers exist between test results and teaching:

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

-

ON, High school English consultant, female

Ongoing assessment guides teaching:

I am trying to think of a way that it [provincial assessment] has directly impacted us here in our physical plant, I just, I don't know. I am failing to come up with anything especially positive for those people who have the boots on the ground for each class. . . I just think the teacher knows before they even write, before the assessment is even written you can, you can forecast, umm, for the most part because you know because of the assessments you do on a daily, weekly, monthly basis. The assessment isn't going to make us know our students better, is another way of putting it. - PEI, K-9 school principal, male

The process of large-scale assessment does not always provide value-for-money:

I don't know of any teachers, and, as I say, I was an administrator for many years so I've dealt with this issue on a regular basis even though I haven't taught the other classes, there are no teachers who think the provincial exams are valuable, a valuable process.

- QC, High school English teacher, male

Good teachers, caring teachers, get the best results:

You want me to be honest [about what improves academic results]? Just good teachers; authentic teachers that . . . cared for the kids and cared about not just those subject areas that were being evaluated but every subject. They were just good teachers, but had a very strong personal and social relationship with their kids, and I think that was very important. I saw some very technical teachers, very technically sound in content areas being assessed, and they were not as effective as teachers that were just . . . that had that relationship piece down with their students. Kids know if they are cared for, and you can teach them.

- SK, Elementary school principal, male (a)

7.10 Charts

Figure 7.20: Residual analysis for background factors and negative reactivity regressions

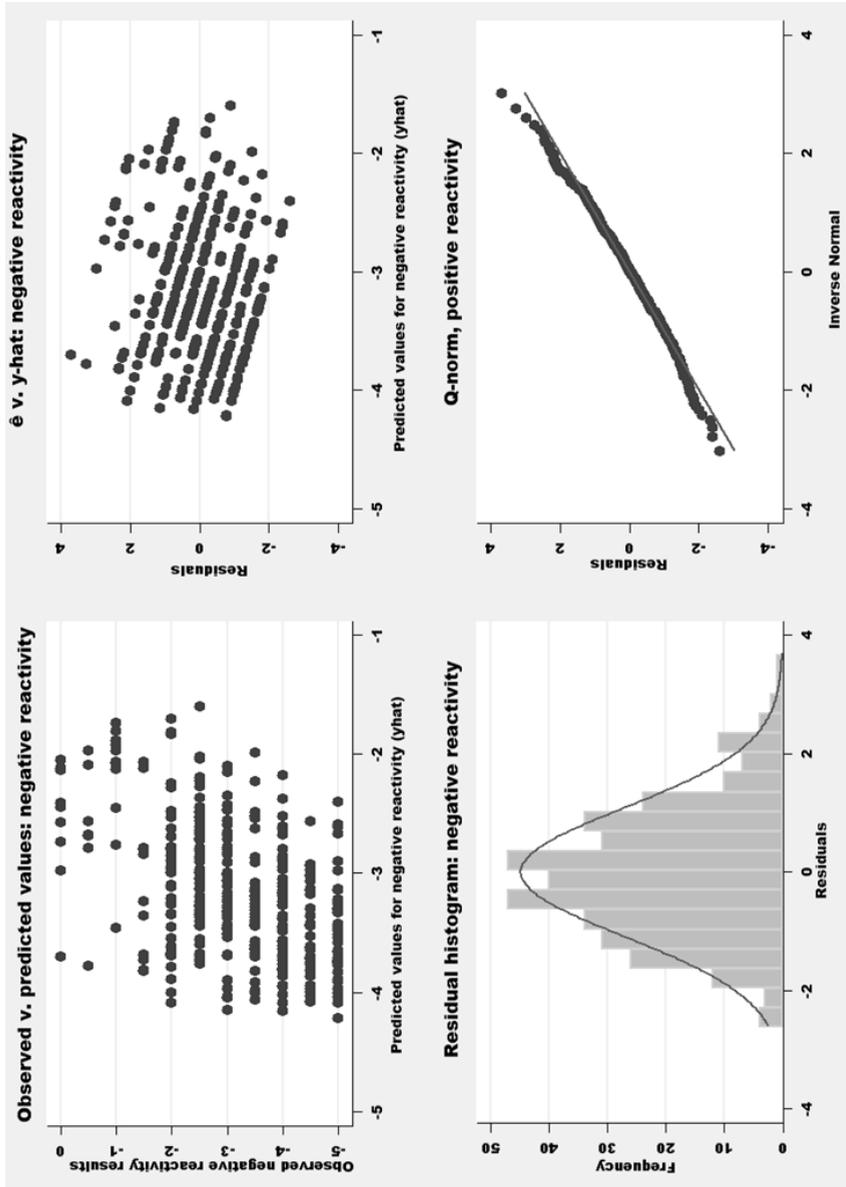
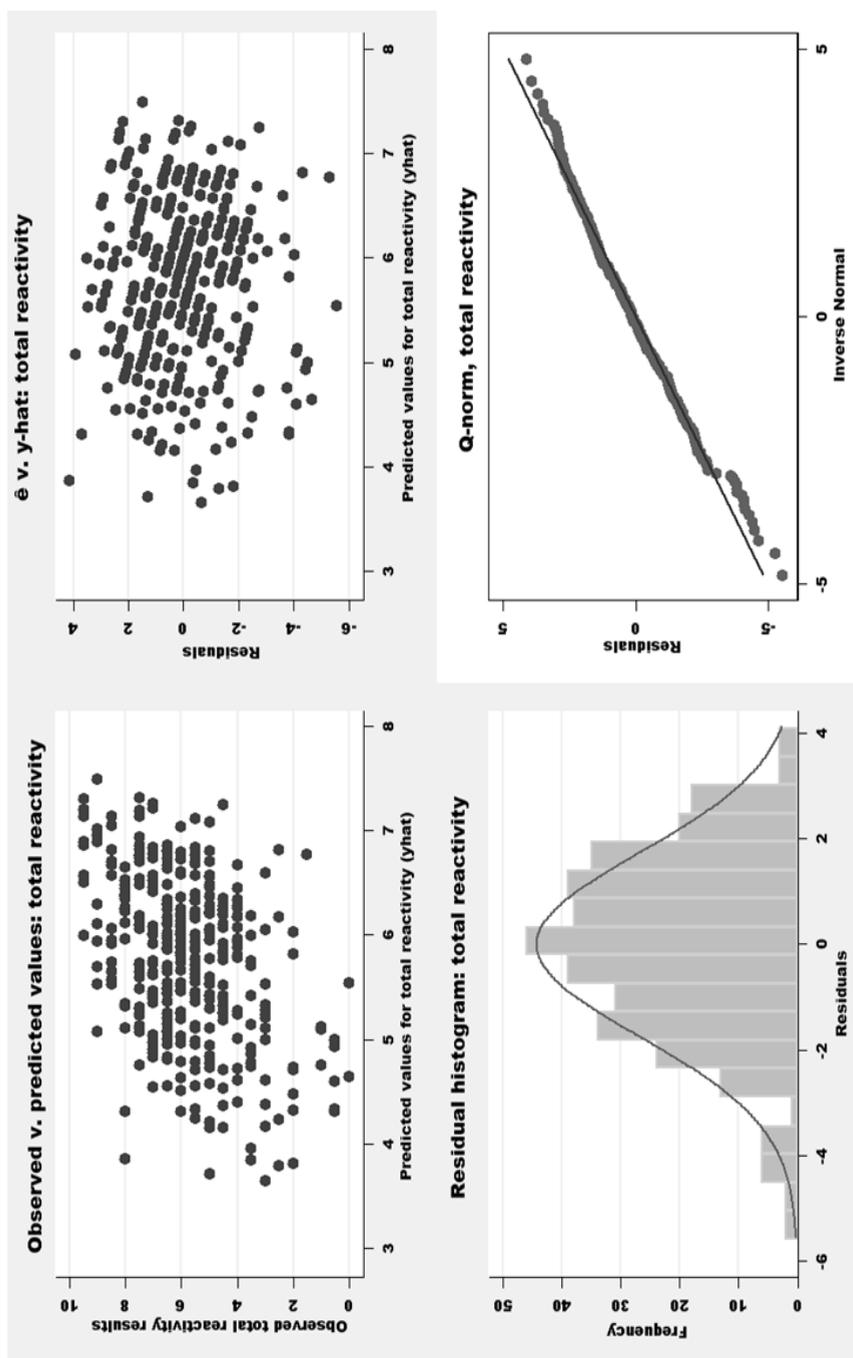


Figure 7.21: Residual analysis for background factors and total reactivity regressions



Summary, conclusions and recommendations

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

- Section 8.1 examines the scale and purpose of the study by looking at the research question and also how and when data were collected
- Section 8.2 presents the reactivity results examined without the inclusion of independent variables
- Section 8.3 shows how test design/data independent variables affect teachers' use of large-scale assessment (LSA) data
- Section 8.4 submits results from attitudes-based independent variables and their influence on reactivity
- Section 8.5 looks at the effects of provided supports on reactivity effects
- Section 8.6 examines the influence of incentives on reactivity effects
- Section 8.7 presents reactivity results in light of background factors such as teacher experience, class size, and school setting
- and Section 8.8 includes recommendations based on the study's findings

8.1 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.

8.2 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.

8.3 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.⁹⁶

Including provincial dummies, these correlations are quite strong, with R² values going from 14% (positive), to 18% (negative) and finally 16% (total).

8.4 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.⁹⁷**
- **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 telling factors were the belief that LSAs can

⁹⁶ See, for example: Lukin, Bandalos, Eckhout & Mickelson, 2004; Hargreaves, Crocker, Davis, McEwen, Sahlberg, Shirley, & Sumara, 2009; Earl & Fullan, 2003.

⁹⁷ Using data for student accountability, for school improvement, and agreement that there are more appropriate uses for the data.

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.

8.5 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.

8.6 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 be 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%).

8.7 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.

8.8 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.

Bibliography

- Abrams, L. M. (2004). Teachers' views on high-stakes testing: Implications for the classroom. *Education Policy Studies Laboratory*. EPSSL Working Paper EPSSL-0401-104-EPRU. Retrieved Mar. 30, 2013 from <http://epsl.asu.edu/epru/documents/EPSSL-0401-104-EPRU.pdf>
- Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, 42(1), 18–29. doi:10.1353/tip.2003.0001
- Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3).
- Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement*, 20(2), 103–118.
- Aitken, N., Webber, C. F., Lupart, J., Scott, S., & Runté, R. (2011). Assessment in Alberta: Six areas of concern. *The Educational Forum*, 75(3), 192–209. doi:10.1080/00131725.2011.576803
- Alberta Education. (n.d.). Provincial achievement tests (PAT). Retrieved Apr. 24, 2013, from <http://education.alberta.ca/admin/testing/achievement.aspx>
- Alexander, E. R., & Faludi, A. (1988). Planning and plan implementation - Notes on evaluation criteria. *Environment and Planning B: Planning and Design*, 16.
- Allen, J. R. (2002). Value-for-money in Saskatchewan K – 12 educational expenditures. *Saskatchewan Institute of Public Policy*. SIPP Public Policy Paper No. 10. Retrieved May 3, 2013 from <http://www.uregina.ca/sipp/documents/pdf/SaskEducationFinal.pdf>
- Amrein, A.L. & Berliner, D.C. (2002, Mar. 28). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18).
- Amrein, A. L., & Berliner, D. C. (2002). An analysis of some unintended and negative consequences of high-stakes testing. *Economic Policy Research Unit*. EPRU Working paper. Retrieved Nov. 26, 2013 from <http://nepc.colorado.edu/files/EPSSL-0211-125-EPRU.pdf>

- Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives*, 18(14), 1–36.
- Amrein-Beardsley, A., Collins, C., Polasky, S. A., & Sloat, E. F. (2013). Value-added model (VAM) research for education policy: Framing the issue. *Education Policy Analysis Archives*, 21(4), 1–14.
- Anderson, G. (1998). Toward authentic participation: Deconstructing the discourses of participatory reforms in education. *American Educational Research Journal*, 35(4), 571-603.
- Anderson, S. E. (2003). The school district role in educational change: A review of the literature. *ICEC Working Paper #2*. Retrieved June 18, 2013 from https://sdcoe.net/lret2/dsi/pdf/District_Role_Change.pdf
- Anderson, S., Leithwood, K., & Strauss, T. (2010). Leading data use in schools : Organizational conditions and practices at the school and district levels. *Leadership and Policy in Schools*, 9(3), 292–327. doi: 10.1080/15700761003731492
- Argyris, C. (2008). Single-loop and double-loop models in research on decision making. *Administrative Science Quarterly*, 21(3), 363–375.
- Armstrong, J., & Anthes, K. (2001). How data can help: Putting information to work to raise student achievement. *American School Board Journal*, (November), 38–41.
- Ashton, P., Buhr, D., & Crocker, L. (1984). Teachers' sense of self-efficacy: A self-referenced or norm-referenced construct? *Florida Journal of Educational Research*, 26(1), 29-41.
- Assael, H., & Keon, J. (1982). Nonsampling vs. sampling errors in survey. *Journal of Marketing*, 46(2), 114–123.
- Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258–267. doi:10.3102/0013189X07306523
- Bacon, A. (1995). The teachers' perspective on accountability. *Canadian Journal of Education*, 20(1), 85-91.

- Baker, E. L. (2007). The end(s) of testing. *Educational Researcher*, 36(6), 309–317. doi:10.3102/0013189X07307970
- Baker, B. D., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, 21(9), 1–71.
- Baker, E., Barton, P. E., Haertel, E., Ladd, H. F., Darling-Hammond, L., Linn, R. L., Ravitch, D., et al. (2010). Problems with the use of student test scores to evaluate teachers. *Economic Policy Institute*. EPI Briefing Paper #278. Retrieved Feb. 01, 2013 from <http://www.epi.org/page/-/pdf/bp278.pdf>
- Baker, E. L., & Linn, R. L. (2002). Validity issues for accountability systems. *Center for the Study of Evaluation*. CSE Technical Report 585. (pp. 1–29). Retrieved Apr. 10, 2013 from <https://www.cse.ucla.edu/products/reports/TR585.pdf>
- Ball, S. J. (1998). Big policies / small world : An introduction to international perspectives in education policy. *Comparative Education*, 34(2), 119–130.
- Bampton, R., & Cowton, C. J. (2002). The E-Interview. *Forum : Qualitative Social Research Sozialforschung*, 3(2).
- Banicky, L. A., & Noble, A. J. (2001). Detours on the road to reform: When standards take a back seat to testing. *Delaware Education Research & Development Center technical report*. Retrieved Mar. 30, 2013 from <http://putnam.lib.udel.edu:8080/dspace/bitstream/handle/19716/2407/t010222.pdf?sequence=1>
- Barton, P. E. (1999). Too much testing of the wrong kind; Too little of the right kind in K-12 education. *Educational Testing Service*. Retrieved Feb. 01, 2013 from <http://www.ets.org/Media/Research/pdf/PICTOOMUCH.pdf>
- Bauer, S. C. (2000). Should achievement tests be used to judge school quality? *Education Policy Analysis Archives*, 8(46), 1–18.
- BC Teachers Federation. (n.d.). *Testing and Assessment*. Retrieved November 26, 2013, from <http://www.bctf.ca/IssuesInEducation.aspx?id=5642>
- Ben Jaafar, S., & Anderson, S. (2007). Policy trends and tensions in accountability for educational management and services in Canada. *The Alberta Journal of Educational Research*, 53(2), 207–227.

- Ben Jaafar, S. & Earl, L. (2008). Comparing performance-based accountability models: A Canadian example. *Canadian Journal of Education*, 31(3), 697–725.
- Benveniste, L. (2000). Student assessment as a political construction : The case of Uruguay. *Education Policy Analysis Archives*, 8(32), 1–40.
- Benveniste, L. (2002). The political structuration of assessment : Negotiating state power and legitimacy. *Comparative Education Review*, 46(1), 89–118.
- Berliner, D. C. (1993). Educational reform in an era of disinformation. *Education Policy Analysis Archives*, 1(2), 1–35.
- Berry, B., Smylie, M., & Fuller, E. (2008). Understanding teacher working conditions : A Review and look to the future. *Center for Teaching Quality*. Retrieved Jan. 10, 2013 from http://www.teachingquality.org/pdfs/TWC2_Nov08.pdf
- Berry, B., Wade, C., & Trantham, P. (2008). Using data, changing teaching. *Educational Leadership*, 66(4), 80-84.
- Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *The American Economic Review*, 87(2), 260–264.
- Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *The Journal of Economic Education*, 29(2), 171–182.
- Bishop, J. H., & Wößmann, L. (2004). Institutional effects in a simple model of educational production. *Education Economics*, 12(1), 17–38.
- Black, P. (2000). Research and the development of educational assessment. *Oxford Review of Education*, 26(3), 407–419. doi:10.1080/3054980020001918
- Black, P., & Wiliam, D. (1989). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139–144.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1). doi:10.1080/0969595980050102
- Blaikie, N. (2000). *Designing social research*. Polity Press: Cambridge.

- Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*(2), 205–225.
- Blok, H., Slegers, P., & Karsten, S. (2008). Looking for a balance between internal and external evaluation of school quality: evaluation of the SVI model. *Journal of Education Policy, 23*(4), 379–395. doi:10.1080/02680930801923773
- Boardman, A. G., & Woodruff, A. L. (2004). Teacher change and “high-stakes” assessment: What happens to professional development? *Teaching and Teacher Education, 20*(6), 545–557. doi:10.1016/j.tate.2004.06.001
- Bolon, C. (2000). School-based standard testing. *Education Policy Analysis Archives, 8*(23), 1–43.
- Booher-Jennings, J. (2013). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268.
- Borghans, L., Meijers, H., & Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry, 46*(1), 2–12. doi:1111/j.1465-7295.2007.00073.x
- Borghans, L., Kockelkorn, L., & Schils, T. (2013). Low stakes, high stakes: The predictive power of math achievement tests. Retrieved Mar. 18, 2013 from http://www.roa.unimaas.nl/seminars/pdf2013/Lex_Borghans.pdf
- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8). doi:10.3102/0013189X033008003
- Borko, H., Elliott, R., & Uchiyama, K. (2002). Professional development: A key to Kentucky's educational reform effort. *Teaching and Teacher Education, 18*(8), 969–987. doi:10.1016/S0742-051X(02)00054-9
- Borman, G. D., & Kimball, S. M. (2004). Teacher quality and educational equality : Do teachers with higher standards-based evaluation ratings close student achievement gaps? *Consortium for Policy Research in Education*. CPRE-UW Working Paper TC-04-03. Retrieved Sept. 16, 2014 from http://www.cpre.wceruw.org/papers/Teacher_Equity_AERA04.pdf

- 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).
- Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). Estimating the effect of leaders on public sector productivity: The case of school principals. *National Bureau of Economic Research*. NBER Working Paper 17803. Retrieved Dec. 2, 2012 from http://www.iza.org/conference_files/Leadership_2012/hanushek_e2570.pdf
- Brannen, J. (2005). Mixing methods: The entry of qualitative and quantitative approaches into the research process. *International Journal of Social Research Methodology, 8*(3), 173–184. doi:10.1080/13645570500154642
- Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1).
- Breakspear, S. (2012). The policy impact of PISA: an exploration of the normative effects of international benchmarking in school system performance. *Organization for Economic Cooperation and Development*. OECD Education Working Papers, 71. doi: [10.1787/5k9fdqffr28-en](https://doi.org/10.1787/5k9fdqffr28-en)
- Bredeson, P. V., & Kose, B. W. (2007). Responding to the education reform agenda: A study of school superintendents' instructional leadership. *Education Policy Analysis Archives, 15*(5).
- Breiter, A., & Light, D. (2006). Data for school improvement: Factors for designing effective information systems to support decision-making in schools. *Journal of Educational Technology and Society, 9*(3), 206–217.
- British Columbia Ministry of Education. (n.d.). Provincial student assessment program. Retrieved Apr. 24, 2013, from <http://www.bced.gov.bc.ca/assessment/>
- Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301–318. doi: /10.1080/0969594042000304609
- Brownlee, W. (1995). Accountability initiatives: Necessary or contrived? *Canadian Journal of Education, 20*(1), 80-84.

- Bryk, A. S., & Hermanson, K. L. (1993). Educational indicator systems : Observations on their structure, interpretation, and use. *Review of Research in Education, 19*, 451–484.
- Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research, 1*(1), 8–22. doi:10.1177/2345678906290531
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4).
- Campbell, D. (1969). Reforms as experiments. *American Psychologist, 24*(4), 409–429.
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.
- Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). Linking provincial student assessments with national and international assessments. *Statistics Canada: Education, Skills and Learning - Research Papers*. Retrieved Apr. 16, 2012, from <http://publications.gc.ca/Collection/Statcan/81-595-MIE/81-595-MIE2003005.pdf>
- Chakwera, E., Khembo, D., & Sireci, S. G. (2004). High-stakes testing in the warm heart of Africa : The challenges and successes of the Malawi National Examinations Board. *Education Policy Analysis Archives, 12*(29), 1–21.
- Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics, 126*(4), 1593–1660. doi:10.1093/qje/qjr041
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. *National Bureau of Economic Research*. NBER Working Paper Series 17699. Retrieved Apr. 14, 2014 from <http://www.nber.org/papers/w17699>
- Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2).

- Cirtwell, C., & O'Keefe, B. (2008, July). One Size Fits None: Putting kids' achievement first, comes with putting kids first. *AIMS - Atlantic Institute for Market Studies*. Retrieved June 6, 2012, from <http://www.aims.ca/en/home/library/details.aspx/2249>
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4).
- Cizek, G. J. (2005). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice*, 19(2), 16–23. doi:10.1111/j.1745-3992.2000.tb00026.x
- Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states. *National Board on Educational Testing and Public Policy*. NBETPP Report. Retrieved May 17, 2013 from <http://www.eric.ed.gov/pdfsed474867.pdf>
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). How and why do teacher credentials matter for student achievement? *National Center for the Analysis of Longitudinal Data in Education Research*, CALDER Working Paper 2. Retrieved Oct. 11, 2014 from <http://www.nber.org/papers/w12828.pdf>
- Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts : Mapping the terrain. *American Journal of Education*, 112(4), 469–495.
- Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, 9(4), 173–206. doi:10.1080/15366367.2011.626729
- Coburn, C. E., & Russell, J. L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis*, 30(3), 203–235. doi:10.3102/0162373708321829
- Cochran, W. G. (1977). *Sampling techniques* (3d ed.). New York: Wiley.
- Cohen, D. K. (1995). What is the system in systemic reform? *Educational Researcher*, 24(9), 11–17.
- Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice : A commentary. *Education Policy Analysis Archives*, 12(3), 331–338.

- Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance : The mathematics reform in California. *Teachers College Record*, 102(2), 294–343. doi:10.1111/0161-4681.00057
- Coleman, James, S. (1967, Oct. 2). The concept of equality of educational opportunity. *US Office of Education Conference*. Retrieved Sep. 16, 2014 from <http://files.eric.ed.gov/fulltext/ED015157.pdf>
- Colombo, R. (2000). A model for diagnosing and reducing nonresponse bias. *Journal of Advertising Research*.
- Corcoran, T. Fuhrman, S. & Belcher, C. (2001). The district role in instructional improvement. *Phi Delta Kappan*, 83(1), 78–84.
- Corcoran, T. & Goertz, M. (1995). Instructional capacity and high performance schools. *Educational Researcher*, 24(9), 08.
- Costrell, R. M. (1994). A simple model of educational standards. *The American Economic Review*, 84(4), 956–971. doi:10.1126/science.151.3712.867-a
- Cotton, K. (1996). School size, school climate, and student performance. *School Improvement Research Series, Close-up #20*. Retrieved Sep. 16, 2014 from March 26, 2003 from https://castl.duq.edu/conferences/Newmiddle/School_Size.pdf
- Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64(4), 464–94.
- Covaleskie, J. F. (1994). The educational system and resistance to reform : The limits of policy. *Education Policy Analysis Archives*, 2(4), 1–10.
- Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. *Urban Education*, 42(6).
- Crundwell, R. M. (2005). Alternative strategies for large scale assessment in Canada: Is value-added assessment one possible answer. *Canadian Journal of Educational Administration and Policy*, 41.

- Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. *National Bureau of Economic Research*. NBER Working Paper Series 12286. Retrieved Apr. 1, 2013 from <http://www.nber.org/papers/w12286>
- Darling-Hammond, L. (1990). Instructional policy into practice: "The power of the bottom over the top". *Educational Evaluation and Policy Analysis*, 12(3), 339–347.
- Darling-Hammond, L. (2003). Standards and assessments: Where we are and what we need. *Teachers College Record*.
- Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, 106(6), 1047–1085.
- Darling-Hammond, L. & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education* 104(2). doi:10.1111/j.1744-7984.2005.00034.x/pdf
- Davidson, K. L., & Frohbieter, G. (2011). District adoption and implementation of interim and benchmark assessments. *National Center for Research on Evaluation, Standards, and Student Testing*. CRESST Report 806. Retrieved Mar. 30, 2013 from <https://www.cse.ucla.edu/products/reports/R806.pdf>
- De Wit, K., & Verhoeven, J. C. (2004). Autonomy vs. control: Quality assurance and governmental policy in Flanders. *Education Policy Analysis Archives*, 12(71).
- Debard, R., & Kubow, P. K. (2002). From compliance to commitment: The need for constituent discourse in implementing testing policy. *Educational Policy*, 16(3), 387–405. doi:10.1177/08904802016003002
- Deci, E. L., Spiegel, N. H., Ryan, R. M., Koestner, R., & Kauffman, M. (1982). Effects of performance standards on teaching styles: Behavior of controlling teachers. *Journal of Educational Psychology*, 74(6), 852–859.
- Dee, T. S., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management*, 23(3), 471–488. doi:10.1002/pam.20022
- Demoss, K. (2002). Leadership styles and high-stakes testing: Principals make a difference. *Education and Urban Society*, 35(1), 111–132. doi:10.1177/001312402237217

- Denzin, N. K., & Lincoln, Y. S. (2000). Introduction: The discipline and practice of qualitative research. In N. K. Denzin & Y.S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 2-32). Thousand Oaks, Calif.: Sage.
- Denzin, N. K. (2010). Moments, mixed methods, and paradigm dialogs. *Qualitative Inquiry*, 16(6), 419–427. doi:10.1177/1077800410364608
- Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research*, 72(3), 433–479.
- Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, 24(2).
- Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1).
- Duemer, L. S., & Mendez-Morse, S. (2002). Recovering policy implementation : Understanding implementation through informal communication. *Education Policy Analysis Archives*, 10(39), 1–11.
- DuFour, R., DuFour, R., Eaker, R., & Many, T. (2006). *Learning by doing: a handbook for professional learning communities at work*. Bloomington, Ind.: Solution Tree.
- Duncan, C. R., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *The Alberta Journal of Educational Research*, 53(1), 1–21.
- Dunleavy, J. (2007). *Public education in Canada: Facts, trends, and attitudes - 2007*. Toronto: Canadian Education Association. Retrieved May 17, 2013 from <http://www.cea-ace.ca/sites/cea-ace.ca/files/cea-2007-public-education-in-canada.pdf>
- Earl, L. M. (1995). Assessment and accountability in education in Ontario. *Canadian Journal of Education*, 20(1), 45-55.
- Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education*, 33(3), 383–394. doi:10.1080/0305764032000122023
- Educational Quality and Accountability Office. (n.d.). EQAO. Retrieved Apr. 24, 2013, from <http://www.eqao.com>

- Ehren, M. C. M., & Swanborn, M. S. L. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 23(2), 257–280.
- Elementary Teachers' Federation of Ontario (2001). *Adjusting the optics: Assessment, evaluation and reporting*. Retrieved Mar. 6, 2014 from [http://www.etfo.ca/Publications/PositionPapers/Documents/Adjusting the Optics - Assessment, Evaluation and Reporting.pdf](http://www.etfo.ca/Publications/PositionPapers/Documents/Adjusting%20the%20Optics%20-%20Assessment,%20Evaluation%20and%20Reporting.pdf)
- Elmore, R. F. (1979). Backward mapping: Implementation research and policy decisions. *Political Science Quarterly*, 94(4).
- Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, 66(1).
- Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113.
- Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*, 15(2), 195–219.
- Fagan, L. P. (1995). Performance accountability in the Newfoundland school system. *Canadian Journal of Education*, 20(1), 65-76.
- Falk, B., & Ort, S. (1998). Sitting down to score: Teacher learning through assessment. *Phi Delta Kappan*, 80(1), 59–64.
- Fehr, D. (2008). Financial industry certification preparation and "teaching to the test". *Journal of Economics and Finance Education*, 7(1).
- Ferrão, M. E. (2012). On the stability of value added indicators. *Quality & Quantity*, 46(2), 627–637. doi:10.1007/s11135-010-9417-6
- Ferraro, F., Pfeffer, J., & Sutton, R. I. (2005). Economics language and assumptions: How theories can become self-fulfilling. *The Academy of Management Review*, 30(1).
- Feuer, M. J., Towne, L., & Shavelson, R. J. (2004). Scientific culture and educational research. *Zeitschrift für Erziehungswissenschaft*, 7(4).

- Figlio, D. N., & Getzler, L. S. (2002). Accountability, ability and disability: Gaming the system. *National Bureau of Economic Research*. NBER Working Paper 9307. Retrieved Apr. 18, 2013 from <http://www.nber.org/papers/w9307>
- 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. doi:10.3102/0002831207306767
- Firestone, W. A. (1989). Using reform: Conceptualizing district initiative. *Educational Evaluation and Policy Analysis*, 11(2), 151–164.
- Firestone, W. A., Mangin, M. M., Martinez, M. C., & Polovsky, T. (2005). Leading coherent professional development: A comparison of three districts. *Educational Administration Quarterly*, 41(3), 413–448. doi:10.1177/0013161X04269605
- Firestone, W. A., Mayrowetz, D., & Fairman, J. (1989). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, 20(2), 95–113.
- Firestone, W. A., Monfils, L., Camilli, G., Schorr, R. Y., Hicks, J. E., & Mayrowetz, D. (2002). The ambiguity of test preparation: A multimethod analysis in one state. *Teachers College Record*, 104(7), 1485–1523. doi:10.1111/1467-9620.00211
- Firestone, W. A., Schorr, R. Y., & Monfils, L. F. (Eds.). (2004). *The ambiguity of teaching to the test: Standards, assessment, and educational reform*. Mahweh, NJ: L. Erlbaum Associates Publishers
- Fitz, J. (1994). Implementation research and education policy : Practice and prospects. *British Journal of Educational Studies*, 42(1), 53–69.
- Flick, U. (2002). Qualitative research - State of the art. *Social Science Information*, 41(1), 5–24. doi:10.1177/0539018402041001001
- Flick, U. (2006). *An introduction to qualitative research* (3rd ed.). London: Sage Publications.
- Florida Department of Education. (n.d.). Bureau of K-12 assessment. Retrieved Apr. 24, 2013 from <http://fcats.fldoe.org/>

- Fountain, J. E. (2001). Paradoxes of public sector customer service. *Governance*, 14(1), 55–73. doi:10.1111/0952-1895.00151
- Fournier, G. (2000). The Pan-Canadian assessments: Setting the record straight. *Phi Delta Kappan*, 81(7), 547–550.
- Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39(3), 193–202.
- Fricker, R. D. (2006). Sampling methods for web and e-mail surveys. Retrieved Mar. 3, 2013 from [http://www.nps.navy.mil/orfacpag/resumePages/papers/frickerpa/Draft Internet Survey Sampling Chapter.pdf](http://www.nps.navy.mil/orfacpag/resumePages/papers/frickerpa/Draft%20Internet%20Survey%20Sampling%20Chapter.pdf)
- Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of internet research surveys: Evidence from the literature. *Field Methods*, 14(4), 347–367. doi:10.1177/152582202237725
- Fuchs, T., & Wößmann, L. (2007). What accounts for international differences in students' performance? A re-examination using PISA data. *Empirical Economics*, 32.
- Fullan, M. (1985). Change processes and strategies at the local level. *The Elementary School Journal*, 85(3), 390–421.
- Fullan, M. (2005). The Tri-level solution: School/district/state synergy. *Education Analyst*, (Winter), 4–5.
- Fullan, M. (2009). Large-scale reform comes of age. *Journal of Educational Change*, 10(2-3), 101–113. doi:10.1007/s10833-009-9108-z
- Fullan, M. (2011). Choosing the wrong drivers for whole system reform. *Centre for Strategic Education Seminar Series*, 204. Retrieved Apr. 13, 2012, from http://www.michaelfullan.ca/home_articles/SeminarPaper204.pdf
- Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research*, 47(2), 335–397.
- Furrer, O., & Sudharshan, D. (2001). Internet marketing research: opportunities and problems. *Qualitative Market Research: An International Journal*, 4(3), 123–129. doi:10.1108/13522750110393026

- Gabriel, R., & Lester, J. N. (2013). Sentinels guarding the grail: Value-added measurement and the quest for education reform. *Education Policy Analysis Archives*, 21(9).
- Gambell, T., & Hunter, D. (2004). Teacher scoring of large-scale assessment: Professional development or debilitation? *Journal of Curriculum Studies*, 36(6). doi:10.1080/0022027032000190696
- Garet, M. S., Porter, A. C., Desimone, L., & Birman, B. F. (2009). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38(4), 915–945.
- Garn, G. A. (1999). Solving the policy implementation problem : The case of Arizona charter schools. *Education Policy Analysis Archives*, 7(26), 1–18.
- Garson, G. D. (2012). *Hierarchical Linear Modeling: Guide and Applications*. SAGE Publications. Retrieved June 18, 2014 from http://www.sagepub.com/upm-data/47528_ch_1.pdf
- Gerring, J. (2012). *Social science methodology: a unified framework* (2nd ed.). Cambridge: Cambridge University Press.
- Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: a construct validation. *Journal of Educational Psychology*, 76(4).
- Goe, L. (2008). Key issue : Using value-added models to identify and support highly effective teachers. *National Comprehensive Center for Teacher Quality*. Retrieved Mar. 31, 2013 from <http://www2.tqsources.org/strategies/het/UsingvalueAddedModels.pdf>
- Goertz, M. E. (2001). Redefining government roles in an era of standards-based reform. *Phi Delta Kappan*, 83(1), 62–66.
- Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). From testing to teaching: The use of interim assessments in classroom instruction. *Consortium for Policy Research in Education (CPRE)*. CPRE Research Report #RR-65. Retrieved Feb.16, 2014 from <http://dx.doi.org/10.1037/e546712012-001>
- Goldhaber, D., & Anthony, E. (2005). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Urban Institute*. Retrieved from <http://files.eric.ed.gov/fulltext/ED490921.pdf>

- Grant, S. G. (2000). Teachers and tests: Exploring teachers' perceptions of changes in the New York State testing program. *Education Policy Analysis Archives*, 8(14), 1–28.
- Graue, M. E., Delaney, K. K., & Karch, A. S. (2013). Ecologies of education quality. *Education Policy Analysis Archives*, 21(9).
- Greene, J. P., Winters, M. A., & Forster, G. (2003). Testing high stakes tests: Can we believe the results of accountability tests? *Center for Civic Innovation*. Retrieved Dec. 19, 2013 from http://heartland.org/sites/all/modules/custom/heartland_migration/files/pdfs/1745.pdf
- Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, 29(1), 158–67.
- Grindle, M. S. (2007). Good Enough Governance Revisited. *Development Policy Review*, 25(5), 533–574. doi:10.1111/j.1467-7679.2007.00385.x
- Guba, E. G., & Lincoln, Y. S. (1994). *Competing paradigms in qualitative research*. In N. K. Denzin & Y.S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: Sage.
- Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, 8(3/4), 381–391. doi:10.1080/135406002100000512
- 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).
- Halverson, R., & Thomas, C. N. (2007). The roles and practices of student services staff as data-driven instructional leaders. *Wisconsin Center for Education Research*. WCER Working Paper No. 2007-1. Retrieved Apr. 17, 2013 from <http://dx.doi.org/10.1080/00131720709335008>
- Hamilton, L. (2003). Chapter 2: Assessment as a policy tool. *Review of Research in Education*, 27(1), 25–68. doi:10.3102/0091732X027001025

- Hamilton, L. S., & Berends, M. (2006). Instructional practices related to standards and assessments. *RAND Education*. RAND Working Paper WR-374-EDU. Retrieved Apr. 24, 2013 from http://www.rand.org.cn/content/dam/rand/pubs/working_papers/2006/RAND_WR374.pdf
- Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: Rand. Retrieved Apr. 15, 2014 from http://www.rand.org/content/dam/rand/pubs/monograph_reports/2002/MR1554.pdf
- Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, 61(2), 280–288.
- Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Education Evaluation and Policy Analysis*, 19(2), 141–164.
- Hanushek, E.A. (2008). Incentives for efficiency and equity in the school system. *Perspektiven der Wirtschaftspolitik*, 9, 5–27. doi:10.1111/j.1468-2516.2008.00272.x
- Hanushek, E.A., & Raymond, M. E. (2002, June). Improving educational quality: How best to evaluate our schools? Lecture from Education in the 21st Century: Meeting the Challenges of a Changing World, Boston, USA. Retrieved Dec. 2, 2012 from [http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BRaymond 2003 Educ21stCent.pdf](http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BRaymond%2003%20Educ21stCent.pdf)
- Hanushek, E.A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(4), 297-327.
- Hanushek, E.A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2), 267–271. doi:10.1257/aer.100.2.267
- Hanushek, E.A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, 4(1), 131–157.

- Hanushek, E.A., Warren, J. R., & Grodsky, E. (2012). Evidence, methodology, test-based accountability, and educational policy: A scholarly exchange between Dr. Eric A. Hanushek and Drs. John Robert Warren and Eric Grodsky. *Educational Policy*, 26(3), 351–368.
- Hargreaves, A., Crocker, R., Davis, B., McEwen, L., Sahlberg, P., Shirley, D., Sumara, D., et al. (2009). The learning mosaic: A multiple perspectives review of the Alberta Initiative for School Improvement (AISI). Alberta: Alberta Education. Retrieved Mar. 19, 2013 from http://education.alberta.ca/media/6412272/learning_mosaic_full_report_2009.pdf#page=25
- Hargreaves, A., & Shirley, D. (2011). The far side of educational reform. *Canadian Teacher's Federation Brief*. Retrieved Mar.19, 2013 from http://www.ctf-fce.ca/publications/Briefs/Report_EducationReform2012_EN_web.pdf
- Harper, G. F., & Maheady, L. (1991). Factors influencing continued implementation of an educational innovation. *Education*, 111(3), 346–357.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 293–329. doi: 10.3102/003465430298487
- Hay-Gibson, N. V. (2009). Interviews via VoIP: Benefits and disadvantages within a PhD study of SMEs. *Library and Information Research*, 33(105).
- Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, 32(4), 385–410. doi:10.1111/j.1468-2958.2006.00281.x
- Heckman, J. J., & Kautz, T. D. (2012). Hard evidence on soft skills. *National Bureau of Economic Research*. NBER Working Paper 18121.
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31. doi:10.1111/j.1745-3992.2009.00151.x
- Herman, J. L., & Baker, E.L. (2005). Making benchmark testing work. *Educational Leadership*, 63(3), 48-54.

- Herman, J. L., & Golan, S. (1991). Effects of standardized testing on teachers and learning: Another look. *Center for the Study of Evaluation*. CSE Technical Report 334. Retrieved Jan. 11, 2014 from <https://www.cse.ucla.edu/products/reports/Tech334.pdf>
- Herman, J. L., & Gribbons, B. (2001). Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation. *Center for the Study of Evaluation*. CSE Technical Report 535. Retrieved Jan. 11, 2014 from <http://www.cse.ucla.edu/products/reports/TR535.pdf>
- Hobbs, S. A., Walle, D. L., & Hammersly, G. A. (1979). Effects of expectancy of outcome on the reactivity of self-monitoring. *Journal of Psychopathology and Behavioral Assessment*, 1(4).
- Hofmann, D. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, 23(6), 723–744. doi:10.1016/S0149-2063(97)90026-X
- Hollingworth, L., Beard, J. J., & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. *Practical Assessment, Research & Evaluation*, 12(18).
- Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research*, 80(4), 476–526. doi:10.3102/0034654310383147
- Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics, & Organization*, 7(January 1991), 24–52.
- Honig, M. I. (2003). Building policy from practice: District central office administrators' roles and capacity for implementing collaborative education policy. *Educational Administration Quarterly*, 39(3), 292–338. doi:10.1177/0013161X03253414
- Honig, M. I. (2004). Where's the “up” in bottom-up reform? *Educational Policy*, 18(4), 527–561. doi:10.1177/0895904804266640

- Honig, M. I., & Coburn, C. (2007). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, 22(4), 578–608. doi:10.1177/0895904807307067
- Hood, C. (2006). Gaming in targetworld: The targets approach to managing British public services. *Public Administration Review*, 66(4). doi:10.1111/j.1540-6210.2006.00612.x
- Horn, C. (2003). High-stakes testing and students: Stopping or perpetuating a cycle of failure? *Theory into Practice*, 42(1).
- Hox, J. (2010). *Multilevel analysis: techniques and applications* (2nd ed.). New York: Routledge.
- Hudson, D., Seah, L.-H., Hite, D., & Haab, T. (2004). Telephone presurveys, self-selection, and non-response bias to mail and Internet surveys in economic research. *Applied Economics Letters*, 11(4), 237–240. doi:10.1080/13504850410001674876
- Ingersoll, R. M. (1996). Teachers' decision-making power and school conflict. *American Sociological Association*, 69(2), 159–176.
- Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. (2010). *Tri-council policy statement: Ethical conduct for research involving humans*. Ottawa: Interagency Secretariat on Research Ethics. Retrieved Nov.13, 2013 from http://www.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf
- Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, 23(2), 99–121.
- Jacob, B. A. (2002, Jun.). Test-based accountability and student achievement gains: Theory and evidence. *Taking Account of Accountability: Assessing Politics and Policy*. Lecture conducted from John F. Kennedy School of Government, Harvard University, Cambridge MA.
- Jacob, B. A. (2004). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89. doi:10.1016/j.jpubeco.2004.08.004
- Jacob, B. A. (2007) "The Challenges of Staffing Urban Schools with Effective Teachers," *The Future of Children* 17(1): 129-153.

- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3), 843–877.
- James, W. (1997). What pragmatism means. In L. Menard (Ed.), *Pragmatism: A reader* (pp. 93-111). New York: Vintage Books.
- Jerald, B. C. D. (2006). The hidden costs of curriculum narrowing. *The Center for Comprehensive School Reform and Improvement Issue Brief*. Retrieved Feb. 16, 2014 from <http://www.centerforcsri.org/files/CenterIssueBriefAug06.pdf>
- Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24(4), 602–611.
- Johnson, J. H. (1997). Data-driven school improvement. *ERIC Digest*, 109(Jan.), 1–5. Retrieved Jan. 29, 2013 from <http://scholarsbank.uoregon.edu/jspui/bitstream/1794/3331/1/digest109.pdf>
- Johnson, P. E., & Chrispeels, J. H. (2010). Linking the central office and its schools for reform. *Educational Administration Quarterly*, 46(5), 738–775. doi:10.1177/0013161X10377346
- Jones, B. D., & Egley, R. J. (2004). Voices from the frontlines : Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives*, 12(39).
- Jones, B. D., & Egley, R. J. (2008). Learning to take tests or learning for understanding? Teachers' beliefs about test-based accountability. *The Educational Forum*, 71(3), 232–248. doi:10.1080/00131720709335008
- Kane, M. (2008). *Errors of measurement, theory, and public policy*. Retrieved Feb. 01, 2013 from <http://www.ets.org/Media/Research/pdf/PICANG12.pdf>
- Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16(4), 91–114. doi:10.1257/089533002320950993
- Kemp, S., & Friesen, D. (2009, Apr.). Student assessment for teaching and learning: Teacher perceptions and practices. *Saskatchewan Instructional Development and Research Unit (SIDRU)*. Retrieved May 30, 2012, from www.mcdowellfoundation.ca/main_mcdowell/projects/research_rep/180_student_assessment_for_teaching_and_learning.pdf

- Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement : Actions, outcomes, and lessons from three urban districts. *American Journal of Education*, 112(4).
- Kersting, N. B., & Chen, M. (2013). Value-added teacher estimates as part of teacher evaluations : Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives*, 21(9).
- Kish, L. (1995). Chapter 1: Introduction. In L. Kish, *Survey Sampling* (pp. 1–34). New York: Wiley Classics. Retrieved Mar. 3, 2013 from http://www.abacpoll.au.edu/subresearch/bf6993/chapter/readings/oct3_readings/pdf/oct3_3.pdf
- Kjær, A.M. (2004). *Governance*. Cambridge: Polity Press.
- Klinger, D. A., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy*, 76.
- Klinger, D. A., & Rogers, W. T. (2003). An investigation of the accuracy of alternative methods of true score estimation in high stakes mixed-format examinations. *The Alberta Journal of Educational Research*, XLIX(1), 83–97.
- Klingner, J. K., Boardman, A. G., & McMaster, K. L. (2013). What does it take to scale up and sustain evidence-based practices? *Exceptional Children*, 79(2), 195–211.
- Knighton, T., & Brassière, P. (2006). Educational outcomes at age 21 associated with reading ability at age 15. *Statistics Canada: Culture, Tourism and the Centre for Education Studies – Research Papers*. Retrieved Apr. 26, 2012, from <http://publications.gc.ca/Collection/Statcan/81-595-MIE/81-595-MIE2006043.pdf>
- Kober, N. (2002). Teaching to the test: The good, the bad, and who's responsible. *TestTalk for Leaders*. Retrieved Jan. 27, 2013 from <http://www.cep-dc.org/displayDocument.cfm?DocumentID=256>
- Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, 28(6), 682–692.

- Kohn, A. (2001). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan*, 82(5).
- Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources*, 37(4).
- Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, 22, 18–26.
doi:10.1111/j.1745-3992.2003.tb00124.x
- Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education*, 104(2). Retrieved Aug. 8, 2012, from <http://cse.ucla.edu/products/reports/r655.pdf>
- Koretz, D. (2008). A measured approach: Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator*, Fall, 18–28.
- Koretz, D. (2009, Dec.). Implications of current policy for educational measurement. *Center for K-12 Assessment & Performance Management*. Lecture conducted from the Center for K-12 Assessment & Performance Management, Princeton, NJ.
- Koretz, D., & Hamilton, L. (2003). Teachers' responses to high-stakes testing and the validity of gains : A pilot study. *Center for the Study of Evaluation*. CSE Technical Report 610. Retrieved Jan. 27, 2013 from <http://www.cse.ucla.edu/products/reports/r610.pdf>
- Koretz, D. & Jennings, J. L. (2010, Feb. 11). The misunderstanding and use of data from educational tests. *The Process of Data Use*. Lecture conducted from The Spencer Foundation, Chicago IL.
- Koretz, D., McCaffrey, D.F., Hamilton, L. (2001). Toward a framework for validating gains under high-stakes conditions. *Center for the Study of Evaluation*. CSE Technical Report 551. Retrieved Feb. 3, 2013 from <http://www.cse.ucla.edu/products/Reports/TR551.pdf>
- Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy*, 18(1), 45–70.
doi:10.1177/0895904803260024

- Lachat, M. A., & Smith, S. (2009). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk*, (Oct. 2012).
- Ladd, H. F. (2001). School-based educational accountability systems: The promise and the pitfalls. *National Tax Journal*, 54(2), 385–400.
- Laitsch, D. (2005). A policymaker's primer on testing and assessment. *Association for Supervision and Curriculum Development*. Retrieved Feb. 1, 2013 from <http://www.ascd.org/publications/newsletters/policy-priorities/jul05/num42/toc.aspx>
- Lather, P. (2004). This is your father's paradigm: Government intrusion and the case of qualitative research in education. *Qualitative Inquiry*.
- Lee, J. (2001). School reform initiatives as balancing acts: Policy variation and educational convergence among Japan, Korea, England and the United States. *Education Policy Analysis Archives*, 9(13), 1–11.
- Lee, C., & Wiliam, D. (2005). Studying changes in the practice of two teachers developing assessment for learning. *Teacher Development*, 9(2), 265–288.
- Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement*, 18(2), 85–96.
- Leithwood, K., & Jantzi, D. (2007). Review of evidence about school size effects: A policy perspective. *Regina School Division #4*, Working Paper. Retrieved from http://www.edu.pe.ca/ESD/pdf/sop_Review_of_Evidence_about_School_Size_Effects.pdf
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125.
- Levin, B. (2001). Conceptualizing the process of education reform from an international perspective. *Education Policy Analysis Archives*, 9(14), 1–17.
- Levin, B. (2010). Governments and education reform: Some lessons from the last 50 years. *Journal of Education Policy*, 25(6), 739–747.
doi:10.1080/02680939.2010.523793

- Levin, B., & Fullan, M. (2008). Learning about System Renewal. *Educational Management Administration & Leadership*, 36(2), 289–303.
doi:10.1177/1741143207087778
- Lewis, S. G., & Naidoo, J. (2004). Whose theory of participation? School governance policy and practice in South Africa. *Contemporary Issues in Comparative Education*, 6(2), 100–112.
- Lieber, J., Butera, G., Hanson, M., Palmer, S., Horn, E., Czaja, C., & Goodman-jansen, G. (2009). Factors that influence the implementation of a new preschool curriculum: Implications for professional development. *Early Education and Development*, 20(3), 456–481. doi:10.1080/10409280802506166
- Linn, R. L. (1998). Assessments and accountability. *Center for the Study of Evaluation*. CSE Technical Report 490.
- Linn, R. L. (2003). Accountability : Responsibility and reasonable expectations. *Educational Researcher*, 32(7), 3–13.
- Lotto, L. S. (1983). Revisiting the role of organizational effectiveness in educational evaluation. *Education Evaluation and Policy Analysis*, 5(3), 367–378.
- Loughran, J. J. (2002). Effective reflective practice: In search of meaning in learning about teaching. *Journal of Teacher Education*, 53(33).
- Louis, K. S., Febey, K., & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. *Educational Evaluation and Policy Analysis*, 27(2), 177-204.
- Luke, A. (2011). Generalising across borders: Policy and the limits of educational science. *Educational Researcher*, 40(8), 367-377.
- Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 23(2). doi:10.1111/j.1745-3992.2004.tb00156.x/pdf
- Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal*, 91(1).
- Lyons, J. E., & Algozzine, B. (2006). Perceptions of the impact of accountability on the role of principals. *Education Policy Analysis Archives*, 14(16).

- Lytton, H., & Pyryt, M. (1998). Predictors of achievement in basic skills: A Canadian effective schools study. *Canadian Journal of Education*, 23(3), 281–301.
- Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65(3).
- Malen, B., Croninger, R., Muncey, D., & Redmond-Jones, D. (2014). Reconstituting schools: “Testing” the “theory of action.” *Educational Evaluation and Policy Analysis*, 24(2), 113–132.
- Manitoba Education. (n.d.). Assessment and evaluation. Retrieved Apr. 24, 2013, from <http://www.edu.gov.mb.ca/k12/assess/index.html>
- Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, 21(2). doi:10.1080/02671520600615638
- Marshall, J., & Steeves, L. (2008). Improving accountability models in public education: Applying logic models of performance management. *Saskatchewan Institute of Public Policy*. Retrieved May 3, 2013 from [http://www.uregina.ca/sipp/documents/pdf/PPP55 Steeves Marshall ONLINE.pdf](http://www.uregina.ca/sipp/documents/pdf/PPP55%20Steeves%20Marshall%20ONLINE.pdf)
- Martens, K., & Niemann, D. (2010). Governance by comparison: How ratings & rankings impact national policy-making in education. *TranState Working Papers*, 139. Retrieved Mar. 15, 2012, from <http://hdl.handle.net/10419/41595>
- Matheson, L. N., Rogers, L. C., Kaskutas, V., & Dakos, M. (2002). Reliability and reactivity of three new functional assessment measures. *Work: A Journal of Prevention, Assessment and Rehabilitation*, 18.
- Matland, R. E. (1995). Synthesizing the implementation literature : The ambiguity-conflict model of policy implementation. *Journal of Public Administration Research and Theory*, 5(2).
- Maynard, A., & Maynard, A. (2010). Translating evidence into practice : Why is it so difficult? *Public Money and Management*, 27(4), 251–256. doi:10.1111/j.1467-9302.2007.00591.x

- McCambridge, J., & Kypri, K. (2011). Can simply answering research questions change behavior? Systematic review and meta analyses of brief alcohol intervention trials. *PLoS ONE*, 6(10).
- McDonnell, L. M. (1994). Assessment policy as persuasion and regulation. *American Journal of Education*, 102(4), 394–420.
- McDonnell, L. M., & Elmore, R. F. (1987). Getting the job done: Alternative policy instruments. *Education Evaluation and Policy Analysis*, 9(2), 133–152.
- McEwen, N. (1995a). Introduction: Accountability in education in Canada. *Canadian Journal of Education*, 20(1), 1-17.
- McEwen, N. (1995b). Educational accountability in Alberta. *Canadian Journal of Education*, 20(1), 27-44.
- McGaghie, W. C., & Thompson, J. A. (2001). America's best medical schools: A critique of the U.S. News & World Report rankings. *Academic Medicine*, 76(10), 985–992.
- McLaughlin, M. W. (1987). Learning From experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, 9(2), 171–178.
- McLaughlin, M. W. (1990). The Rand change agent study revisited: Macro perspectives and micro realities. *Educational Researcher*, 19(9), 11–16.
- McMillan, J. H. (1997). Understanding and improving teachers' classroom assessment decision making : Implications for theory and practice. *Educational Measurement: Issues and Practice*, 22(4). doi:10.1111/j.1745-3992.2003.tb00142.x
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20(1), 20–32. doi:10.1111/j.1745-3992.2001.tb00055.x
- Means, B., Padilla, C., DeBargar, A., & Bakia, M. (2009). Implementing data-informed decision making in schools—Teacher access, supports and use. *U.S. Department of Education, Office of Planning, Evaluation and Policy Development*. Retrieved from <http://files.eric.ed.gov/fulltext/ED504191.pdf>

- Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives*, 6(13).
- Merton, R. K., & Kendall, Patricia L. (1946). The focused interview. *American Journal of Sociology*, 51(6), 541–557.
- Mintrop, H. (2003). The limits of sanctions in low-performing schools : A study of Maryland and Kentucky schools on probation. *Education Policy Analysis Archives*, 11(3), 1–30.
- Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement - and why we may retain it anyway. *Educational Researcher*, 38(5), 353-364.
- Mitchell, M.N. (2010). Data management using Stata: A practical handbook. College Station: Stata Press.
- Møller, J. (2008). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational Change*, 10(1), 37–46. doi:10.1007/s10833-008-9078-6
- Moran-Ellis, J. (2006). Triangulation and integration: Processes, claims and implications. *Qualitative Research*, 6(1), 45–59. doi:10.1177/1468794106058870
- Morgan, C. (2009, May 27). Transnational governance: the case of the OECD PISA. *Canadian Political Science Association*. Lecture conducted from the CPSA, Ottawa.
- Morgan, S. L., & Taylor Poppe, E. S. (2012). The consequences of international comparisons for public support of K-12 education: Evidence from a national survey experiment. *Educational Researcher*, 41(7).
- Morris, A. (2011). Student standardised testing: Current practices in OECD countries and a literature review. *OECD Education Working Papers*, 65. doi:10.1787/5kg3rp9qbnr6-en
- Mu, M., & Childs, R. (2005). What parents know and believe about large-scale assessment. *Canadian Journal of Educational Administration and Policy*, 37.
- Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, 12(1), 53–74. doi:10.1080/13803610500392236

- Muijs, D., & Harris, A. (2006). Teacher led school improvement: Teacher leadership in the UK. *Teaching and Teacher Education*, 22(8), 961–972. doi:10.1016/j.tate.2006.04.010
- Munn, P. (1991). School boards, accountability and control. *British Journal of Educational Studies*, 39(2), 173–189.
- Muriel, A., & Smith, J. (2011). On educational performance measures. *Fiscal Studies*, 32(2), 187–206. doi:10.1111/j.1475-5890.2011.00132.x
- Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28(4), 338–354.
- Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education*, 25(4).
- Najam, A. (1995). Learning from the literature on policy implementation: A synthesis perspective. *International Institute for Applied Systems Analysis. IIASA Working Paper WP-95-61*. Retrieved Jan. 16, 2013 from <http://webarchive.iiasa.ac.at/Admin/PUB/Documents/WP-95-061.pdf>
- Nardi, P. M. (2003). *Doing survey research: A guide to quantitative methods*. Boston: Pearson Education.
- Nettles, S. M., & Herrington, C. (2007). Revisiting the importance of the direct effects of school leadership on student achievement : The implications for school improvement policy. *Peabody Journal of Education*, 82(4), 724–736.
- New Brunswick Education. (n.d.). K-12 Anglophone sector. Retrieved Apr. 24, 2013, from <http://www.gnb.ca/0000/anglophone-e.asp>
- Newfoundland and Labrador Department of Education. (n.d.). Provincial assessments. Retrieved April 23, 2013, from www.ed.gov.nl.ca/edu/k12/evaluation/crts/index.html
- Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement : Updated analyses with NAEP data. *Education Policy Analysis Archives*, 20(20).

- Noblit, G.W., Eaker, D.J. (Authors) (1987, Apr. 20). Evaluation designs as political strategies. *American Educational Research Association*. Lecture conducted from the American Educational Research Association, Washington DC.
- Noell, G. H., Witt, J. C., Gilbertson, D. N., Ranier, D. D., & Freeland, J. T. (1997). Increasing teacher intervention implementation in general education settings through consultation and performance feedback. *School Psychology Quarterly*, 12.
- Noonan, B., & Renihan, P. (2006). Demystifying assessment leadership. *Canadian Journal of Educational Administration and Policy*, 56. =
- Nova Scotia Ministry of Education and Early Childhood Development. (n.d.). Nova Scotia assessments. Retrieved Apr. 24, 2013, from <http://plans.ednet.ns.ca/nova-scotia-assessments>
- Nuland, S., & Poisson, M. (2009). *Teacher codes: learning from experience*. Paris: International Institute for Educational Planning.
- O'Day, J. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, 72(3), 293–329.
- OECD (2011). *Education at a Glance 2011: OECD Indicators*. doi:10.1787/eag-2011-en
- Oláh, L. N., Lawrence, N., & Riggan, M. (2008, Mar. 27). Learning to learn from benchmark assessment data: How teachers analyze results. *AERA Annual Conference*. Lecture conducted from the American Educational Research Association, New York, USA.. Retrieved Feb. 16, 2014 from http://cpre.org/sites/default/files/meetingpaper/1019_aera2008olahlawrenceriggan.pdf
- Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, 70(5), 737–758. doi:/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. doi:10.1080/13645570500402447
- O'Toole, L. J. (2000). Research on policy implementation: Assessment and prospects. *Journal of Public Administration Research and Theory*, 10(1), 263–288.

- Palys, T., & Atchison, C. (2012). Qualitative research in the digital era: Obstacles and opportunities. *International Journal of Qualitative Methods*, 11(4).
- Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1), 163–193. doi:10.3102/0002831210362589
- Parr, J. M., & Timperley, H. S. (2008). Teachers, schools and using evidence: Considerations of preparedness. *Assessment in Education: Principles, Policy & Practice*, 15(1), 57–71. doi:10.1080/09695940701876151
- Parsons, J., & Beauchamp, L. (n.d.). Action research : The Alberta Initiative for School Improvement (AIS) and its implications for teacher education. *ActionResearch*, 3(3), 120–131.
- Pearl, D. K., & Fairley, D. (1985). Testing for the potential for nonresponse bias in sample surveys. *The Public Opinion Quarterly*, 49(4), 553–560.
- Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. Boston: *National Board on Educational Testing and Public Policy*. Retrieved Apr. 22, 2013 from <http://www.eric.ed.gov/PDFS/ED481836.pdf>
- Pittman, T. S., Emery, J., & Bo, A. K. (1982). Intrinsic and extrinsic motivational orientations: Reward-induced changes in preference for complexity. *Journal of Personality and Social Psychology*, 42(5), 789–797.
- Polikoff, M. S. (2012). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis*, 34(3), 278–294. doi:10.3102/0162373711431302
- Pollock, P. H. (2011). *A Stata companion to political analysis* (2nd ed.). Washington, D.C.: CQ Press.
- Pomplun, M. (2009). State assessment and instructional change : A path model analysis. *Applied Measurement in Education*, 10(3), 217–234. doi:10.1207/s15324818ame1003
- Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, 56, 8–15.

- Popham, W.J. (2001). Teaching to the test. *Educational Leadership*, 58(6), 16-20.
- Popper, K. R. (1959). A survey of some fundamental problems. In K.R. Popper, *The logic of scientific discovery* (pp. 27-34). New York: Basic Books.
- Power, M. (2000). The audit society - Second thoughts. *International Journal of Auditing*, 4(1), 111-119.
- Pressman, J. L., & Wildavsky, A. B. (1973). *Implementation: How great expectations in Washington are dashed in Oakland: or, why it's amazing that Federal programs work at all, this being a saga of the Economic Development Administration as told by two sympathetic observers who seek to build morals* (2nd ed.). Berkeley, Calif.: University of California Press.
- Prince Edward Island Department of Education and Early Childhood development. (n.d.). Provincial assessment program. Retrieved Apr. 24, 2013, from <http://www.gov.pe.ca/eecd/studentassessment>
- Propper, C., & Wilson, D. (2003). The use and usefulness of performance measures in the public sector. *CMPO*. CMPO Working Paper Series No. 03/073. Retrieved Nov. 26, 2013 from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.448&rep=rep1&type=pdf>
- Puk, T. (1999). Formula for success according to TIMSS or the subliminal decay of jurisdictional educultural integrity? *The Alberta Journal of Educational Research*, XLV(3), 225-238.
- Pullin, D. (2013). Legal issues in the use of student test scores and value-added models (VAM) to determine educational quality. *Education Policy Analysis Archives*, 21(9), 1-27.
- Québec Ministère de l'Éducation, du Loisir et du Sport (n.d.). Chapter 6: Evaluation and certification. Retrieved Apr. 24, 2013, from <http://www.mels.gouv.qc.ca/reforme/curricu/anglais/chap06.htm>
- Ranson, S. (2003). Public accountability in the age of neo-liberal governance. *Journal of Education Policy*, 18(5), 459-480. doi:10.1080/0268093032000124848

- Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioural Statistics*, 29(1), 121–129.
- Ravitch, D. (1993). Launching a revolution in standards and assessments. *Phi Delta Kappan*, 74(10), 767–772.
- Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.
- Reichardt, R., Snow, R., Schlang, J., & Hupfeld, K. (2008). Overwhelmed and out: Principals, district policy and teacher retention. *Connecticut Center for School Change*. Retrieved Apr. 15, 2014 from <http://www.nctq.org/nctq/research/1220022778926.pdf>
- Rex, L. a., & Nelson, M. C. (2004). How teachers' professional identities position high-stakes test preparation in their classrooms. *Teachers College Record*, 106(6), 1288–1331. doi:10.1111/j.1467-9620.2004.00380.x
- Rhoades, K., & Madaus, G. (2003). Errors in standardized tests: A systemic problem. *National Board on Educational Testing and Public Policy Monograph*. Retrieved Apr. 16, 2012, from www.bc.edu/research/nbetpp/statements/M1N4.pdf
- Riddell, W. C. (2006). The impact of education on economic and social outcomes: An overview of recent advances in economics. *Canadian Policy Research Networks*. Retrieved June 4, 2013 from http://www.rcrpp.org/documents/44362_fr.pdf
- Ritzen, J. (2013). *International Large-Scale Assessments as Change Agents* (pp. 13–24). In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research* Dordrecht: Springer. Retrieved Dec. 2, 2012 from http://link.springer.com/chapter/10.1007/978-94-007-4629-9_2
- Rivkin, S. G., Hanushek, E. A., & Kain, J. E. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.
- Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, 45(2). doi:10.1002/pits.20282/pdf

- Rockoff, J. E. (2003). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2), 247–252.
- Roehrig, G. H., Kruse, R. A., & Kern, A. (2007). Teacher and school characteristics and their influence on curriculum implementation. *Journal of Research in Science Teaching*, 44(7), 883–907.
- Rogers, W. T. (2003). Principles for fair student assessment practices for education in Canada. *Journal of Educational Research and Measurement*, 1(1).
- Rogers, W. T., & Ricker, K. L. (2006). Establishing performance standards and setting cut-scores. *The Alberta Journal of Educational Research*, 52(1), 16–24.
- Rorrer, A. K., & Skrla, L. (2005). Leaders as policy mediators: The reconceptualization of accountability. *Theory Into Practice*, 44(1), 53–62.
- Rorrer, A. K., Skrla, L., & Scheurich, J. J. (2008). Districts as institutional actors in educational reform. *Educational Administration Quarterly*, 44(3), 307–357. doi:10.1177/0013161X08318962
- Rosenkvist, M. A. (2010). Using student test results for accountability and improvement: A literature review. *OECD. OECD Education Working Papers*, No. 54. doi:10.1787/5km4htwzvbv30-en
- Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, 11(24), 1–8.
- Ross, J. A., & Gray, P. (2008). Alignment of scores on large-scale assessments and report-card grades. *The Alberta Journal of Educational Research*, 54(3), 327–341.
- Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioural Statistics*, 29(1), 103–116.
- Runté, R. (1998). The impact of centralized examinations on teacher professionalism. *Canadian Journal of Education*, 23(2).
- Rusch, F. R., Menchetti, B. M., Crouch, K., Riva, M., Morgan, T. K., & Agran, M. (1984). Competitive employment: Assessing employee reactivity to naturalistic observation. *Applied Research in Mental Retardation*, 5(3).

- Ryan, T. G., & Joong, P. (2005). Teachers' and students' perceptions of the nature and impact of large-scale reform. *Canadian Journal of Educational Administration and Policy*, 38.
- Sabatier, P. A. (1986). Top-down and bottom-up approaches to implementation research: A critical analysis and suggested synthesis. *Journal of Public Policy*, 6(1), 21–48.
- Sahlberg, P. (2006). Education reform for raising economic competitiveness. *Journal of Educational Change*, 7(4), 259–287. doi:10.1007/s10833-005-4884-6
- Sahlberg, P. (2008). Rethinking accountability in a knowledge society. *Journal of Educational Change*, 11(1), 45–61. doi:10.1007/s10833-008-9098-2
- Sanders, W. L., & Horn, S. P. (1995). Educational assessment reassessed : The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Education Policy Analysis Archives*, 3(6), 1–15.
- Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement. *The University of Tennessee Value-Added Research and Assessment Center (UT-VARAC)*. Retrieved Sep. 16, 2014 from http://www.cgp.upenn.edu/pdf/Sanders_Rivers-TVASS_teacher_effects.pdf
- 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>
- Sapsford, R. J. (2007). *Survey Research* (2nd ed.). London: Sage Publications.
- Saskatchewan Ministry of Education (n.d.) Assessment for learning program (AFL). Retrieved Apr. 24, 2013, from <http://www.education.gov.sk.ca/afl/english/>
- Saskatchewan Ministry of Education. (2010, Feb.). Provincial panel on student achievement. Retrieved May 30, 2012, from www.education.gov.sk.ca/provpanel-student-achieve-report2010
- Saskatchewan Ministry of Education. (2012a, May). Student achievement initiative: Questions and answers. Retrieved May 30, 2012, from www.education.gov.sk.ca/student-achievement-initiative-Q-and-A

- Saskatchewan Ministry of Education. (2012b, May). Student achievement initiative: Background information. Retrieved May 30, 2012, from www.education.gov.sk.ca/student-achievement-annoucement-backgrounder
- Saskatchewan Teachers' Federation (n.d.). STF Code of Professional Competence. Retrieved Sept. 1, 2012, from <https://www.stf.sk.ca/portal.jsp?Sy3uQUnbK9L2RmSZs02CjVy0w7ZkI/ks6g2u00gzAatsk=F#portal.jsp?S3ua0P4leiBvLe5BSdsr0vZGZJmzTYKNX8t/KNvKOzGyZacpsswpYUA==F>
- Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, 74(1), 63–82.
- Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponsive bias in web and paper surveys. *Research in Higher Education*, 44(4), 409–433.
- Saxon, D., Garratt, D., Gilroy, P., & Cairns, C. (2003). Collecting data in the Information Age: Exploring web-based survey methods in educational research. *Research in Education*, 69, 51–66.
- Scafidi, B., Freeman, C., & Dejarnett, S. (2001). Local flexibility within an accountability system. *Education Policy Analysis Archives*, 9(44), 1–24. Retrieved Feb. 16, 2013 from <http://epaa.asu.edu/ojs/article/view/373/499>
- Schein, E. H. (1973). Organizational culture. *American Psychologist*, 45(2), 109–119.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496. doi:10.1016/j.tate.2009.06.007
- Schochet, P. Z., & Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. *National Center for Education Evaluation*. NCEE 2010-4004. Retrieved Apr. 15, 2014 from <http://files.eric.ed.gov/fulltext/ED511026.pdf>
- Schorr, R. Y., Firestone, W. A., & Monfils, L. (2003). State testing and mathematics teaching in New Jersey: The effects of a test without other supports. *Journal for Research in Mathematics Education*, 34(5).

- Scott, S., Webber, C. F., Aitkin, N., & Lupart, J. (2011). Developing teachers' knowledge, beliefs, and expertise: Findings from the Alberta student assessment study. *The Educational Forum*, 75(2).
doi:10.1080/00131725.2011.552594
- Scriffiny, P. L. (2008). Seven reasons for standards-based grading. *Educational Leadership*, 66(2), 70-74.
- Sedransk, J. (1965). Analytical surveys with cluster sampling. *Journal of the Royal Statistical Society*, 27(2), 264-278.
- Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8), 1438-1457.
doi:10.1287/mnsc.1110.1509
- Sharkey, N. S., & Murnane, R. J. (2006) Tough choices in designing a formative assessment system. *American Journal of Education* 112(4).
- Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher*, 32(1), 25-28.
- Shawer, S. F. (2010). Classroom-level curriculum development: EFL teachers as curriculum-developers, curriculum-makers and curriculum-transmitters. *Teaching and Teacher Education*, 26(2), 173-184. doi:10.1016/j.tate.2009.03.015
- Sheldon, K. M., & Biddle, B. J. (1998). Standards, accountability, and school reform: Perils and pitfalls. *Teachers College Record*, 100(1), 164-180.
- Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, 73(3).
- Shepard, L. A. (1990). "Inflated test score gains": Is it old norms or teaching to the test? *Center for Research on Evaluation, Standards and Student Testing*. CSE Technical Report 307. Retrieved Apr. 15, 2014 from <https://www.cse.ucla.edu/products/reports/TR307.pdf>
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-24. doi:10.1111/j.1745-3992.1997.tb00585.x
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7).

- Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, 85(2). doi:10.1080/01619561003708445
- Shepard, L. A., Davidson, K. L., & Bowman, R. (2011). How middle school mathematics teachers use interim and benchmark assessment data. *National Center for Research on Evaluation, Standards, and Student Testing*. CRESST Report 807. Retrieved Mar. 15, 2012, from <http://www.cse.ucla.edu/products/reports/R807.pdf>
- Shepard, L. A., & Dougherty, K. C. (1991, April). Effects of high-stakes testing on instruction. *AERA Annual Conference*. Lecture conducted from the American Educational Research Association, Chicago, Ill. USA. Retrieved Apr. 15, 2014 from <http://www.colorado.edu/education/sites/default/files/attached-files/Effects%20of%20High-Stakes%20Testing.pdf>
- Shore, C. (2008). Audit culture and illiberal governance: Universities and the politics of accountability. *Anthropological Theory*, 8(3), 278–298. doi:10.1177/1463499608093815
- Shore, C., & Wright, S. (2000). *Coercive accountability: The rise of audit culture in higher education*. In M. Stratham (Ed.), *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. New York: Taylor and Francis. Retrieved June 6, 2012, from http://antropologias.descentro.org/files/downloads/2011/07/Strathern_org_Audit_Culture.pdf#page=70
- Sijtsma, K. (2008, November). *On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha*. Retrieved March 2, 2015 from <http://link.springer.com/article/10.1007/s11336-008-9101-0/fulltext.html>
- 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
- Sirotnik, K. A. (2002). Promoting responsible accountability in schools and education. *Phi Delta Kappan*, 83(9), 662–673.

- Sirotnik, K. A. (2004). *Holding accountability accountable: What ought to matter in public education*. New York: Teachers College Press.
- Sirotnik, K. A., & Kimball, K. (1999). Standards for standards-based accountability systems. *Phi Delta Kappan*, 81(3), 209–214.
- Skwarchuk, S.-L. (2004). Teachers' attitudes toward government- mandated provincial testing in Manitoba. *The Alberta Journal of Educational Research*, 50(3), 252–282.
- Smith, C., Hofer, J., Gillespie, M., Solomon, M., & Rowe, K. (2003). How teachers change: A study of professional development in adult education. *National Center for the Study of Adult Learning and Literacy*. NCSALL Reports #25. Retrieved Apr. 15, 2014 from <http://www.ncsall.net/fileadmin/resources/research/report25.pdf>
- Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5).
- Smith, T. M., & Rowley, K. J. (2005). Enhancing commitment or tightening control: The function of teacher professional development in an era of accountability. *Educational Policy*, 19(1), 126–154. doi:10.1177/0895904804270773
- Smith, T., Desimone, L., & Ueno, K. (2005). "Highly qualified" to do what? The relationship between NCLB teacher quality mandates and the use of reform-oriented instruction in middle school mathematics. *Educational Evaluation and Policy Analysis*, 27(1), 75-109.
- Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies*, 31(2), 143–175. doi:10.1080/002202799183205
- Spillane, J. P. (2005). Distributed leadership. *The Educational Forum*, 69.
- Spillane, J. P. (2012). Data in practice : Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, 118 (February), 1–30. Retrieved Apr. 15, 2014 from http://www.distributedleadership.org/DLS/Publications_files/Data Use manuscript 121511.pdf

- Spillane, J. P., Diamond, J. B., Burch, P., Hallett, T., Jita, L., & Zoltners, J. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy*, 16(5), 731–762. doi:10.1177/089590402237311
- Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition : Reframing and refocusing implementation research. *Review of Educational Research*, 72(3), 387–431.
- Spradley, J. P. (1979). *The ethnographic interview*. Belmont: Wadsworth, Thomson Learning.
- Squires, D., Canney, G. F., & Trevisan, M. S. (2004). There is another way: The faculty-developed Idaho Comprehensive Literacy Assessment for K-8 pre-service teachers. *Education Policy Analysis Archives*, 12(62).
- Stang, A., & Jöckel, K. H. (2004). Studies with low response proportions may be less biased than studies with high response proportions. *American Journal of Epidemiology*, 159(2), 204–210. doi:10.1093/aje/kwh009
- Statistics Canada. (2003). *Education indicators in Canada: Report of the Pan-Canadian education indicators program*. Retrieved Apr. 22, 2013 from <http://www.cesc.ca/pceip/PCEIP2003en.pdf>
- 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
- Stecher, B. M., & Hamilton, L. S. (2006). Using test-score data in the classroom. *RAND Education*. Working Paper WR-375-EDU. Retrieved Apr.14, 2014 from http://130.154.3.8/content/dam/rand/pubs/working_papers/2006/RAND_WR375.pdf
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, 42(8), 795–819. doi:10.1002/pits.20113
- Steenbergen, M. R. ., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science*, 46(1), 218–237.

- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10).
- Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), 271–286.
- Stock, J. H., & Watson, M. W. (2007). Introduction to econometrics (2nd ed.). Boston: Pearson/Addison Wesley.
- Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, 36(4), 187–198. doi:10.3102/0013189X07303396
- Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *The American Economic Review*, 67(4), 639–652.
- Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, 10(2-3), 211–227.
- Supovitz, J. (2013, Apr. 28). How teachers use data to facilitate the learning of all students. *AERA Annual Conference*. Lecture conducted from the American Educational Research Association, San Francisco, USA. Retrieved Feb. 16, 2014 from http://cpre.org/sites/default/files/workingpapers/1539_linkingstudy-supovitzaera2013.pdf
- Sutton, R. E. (2004). Teaching under high-stakes testing: Dilemmas and decisions of a teacher educator. *Journal of Teacher Education*, 55(5).
- Sykes, G. (1990). Organizing policy into practice: Reactions to the cases. *Educational Evaluation and Policy Analysis*, 12(3), 349–353.
- Sykes, G. (1996). Reform of and as professional development. *Phi Delta Kappan*, 77(7), 464–467.
- Taras, M. (2010). Using assessment for learning and learning from assessment. *Assessment & Evaluation in Higher Education*, 27(6), 501–510. doi:10.1080/0260293022000020273
- 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/>

- Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2003). A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost. *Center for the Study of Evaluation*. CSE Technical Report 588. Retrieved Nov. 5, 2013 from <http://www.cse.ucla.edu/products/Reports/TR588.pdf>
- Thomas, D. R. (2006). A general inductive approach for analyzing general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2). doi:10.1177/10982
- Thomas, P. (2004). Performance measurement, reporting and accountability: Recent trends and future directions. *Saskatchewan Institute of Public Policy*, SIPP Public Policy Paper No. 23. Retrieved May 3, 2013 from http://www.uregina.ca/sipp/documents/pdf/PPP23_P_Thomas.pdf
- Thomas, P. G. (2007, May). Why is performance-based accountability so popular in theory and difficult in practice? Lecture conducted from the World Summit on Public Governance. (pp. 1–24). Taipei City, Taiwan. Retrieved Apr. 10, 2013 from https://www.ipac.ca/documents/WHY_IS_PERFORMANCE1.pdf
- Travis, J. E. (1996). Meaningful assessment. *The Clearing House*, 69(5), 308–312.
- Trouteaud, A. R. (2004). How you ask counts: A test of internet-related components of response rates to a web-based survey. *Social Science Computer Review*, 22(3), 385–392. doi:10.1177/0894439304265650
- Trujillo, T. M. (2012). The politics of district instructional policy formation: Compromising equity and rigor. *Educational Policy*, 27(3), 531–559. doi:10.1177/0895904812454000
- Tschannen-Moran, M., & Hoy, A. W. (2002, Apr. 2). The influence of resources and support on teachers' efficacy beliefs. *AERA Conference*. Lecture conducted from the American Educational Research Association, New Orleans, USA. Retrieved Nov.5, 2013 from <http://anitawoolfolkhoy.com/pdfs/aera-2002-megan.pdf>
- Uljens, M. (2007, Nov. 22). The hidden curriculum of PISA: the promotion of neo-liberal policy by educational assessment. *FERA Congress*. Lecture conducted from the Finnish Educational Research Association, Vaasa, Finland.
- Ungerleider, C. (2003). Large-scale student assessment: Guidelines for policymakers. *International Journal of Testing*, 3(2).

- Ungerleider, C. (2006). Reflections on the use of large-scale student assessments for improving student success. *Canadian Journal of Education*, 29(3).
- Valli, L., Croninger, R. G., & Walters, K. (2007). Who (else) is the teacher? Cautionary notes on teacher accountability systems. *American Journal of Education*, 113(4), 635–662.
- Van Thiel, S., & Leeuw, F. L. (2002). The performance paradox in the public sector. *Public Performance & Management Review*, 25(3), 267–281.
- Vernaza, N. (2009). *Teachers' responses to high-stakes accountability in Title I elementary schools: A mixed methods study*. Retrieved Apr. 29, 2012 from http://www.broward.k12.fl.us/research_evaluation/researchresults/528Vernaza/528Vernaza.pdf
- Vicente, P., & Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review*, 28(2), 251–267.
doi:10.1177/0894439309340751
- Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, 35.
- Volante, L. (2005). Accountability, student assessment, and the need for a comprehensive approach. *International Journal for Leadership in Learning*, 9.
- Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of Teaching and Learning*, 4(1).
- Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy*, 58.
- Volante, L. (2010). Understanding the connection between large-scale assessment and school improvement planning. *Canadian Journal of Educational Administration and Policy*, (115), 1–26.
- Volante, L. (2013). Canadian policy responses to international comparison testing. *Interchange*, 44(3-4), 169-178.
- Volante, L., & Ben Jaafar, S. (2008). Profiles of education assessment systems worldwide. *Assessment in Education: Principles, Policy & Practice*, 15(2), 201-210.

- Volante, L., & Cherubini, L. (2007). Connecting educational leadership with multi-level assessment reform. *International Journal for Leadership in Learning*, 11(12).
- Volante, L., & Cherubini, L. (2010). Understanding the connections between large-scale assessment and school improvement planning. *Canadian Journal of Educational Administration and Policy*, 115.
- Volante, L., Cherubini, L., & Drake, S. (2008). Examining factors that influence school administrators' responses to large scale assessment. *Canadian Journal of Educational Administration and Policy*, 84.
- Watson, S., & Supovitz, J. (2001). Autonomy and accountability in the context of standards-based reform. *Education Policy Analysis Archives*, 9(32), 1–21.
- 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).
- Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the data-informed district. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 23(2), 159–178.
- Wayman, J. C., Spring, S. D., Lemke, M. A., & Lehr, M. D. (2012). *Using data to inform practice : Effective principal leadership strategies*. Presented at the 2012 AERA Annual Meeting. Retrieved from http://edadmin.edb.utexas.edu/datause/papers/Wayman_Spring_Lemke_Lehr_Principal_Data_Use_Strategies.pdf
- Wayman, J. C., & Stringfield, S. (2006). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education*, 112(Aug.).
- Webb, P. T. (2006). The choreography of accountability. *Journal of Education Policy*, 21(2), 201–214.
- Webber, C. F., Aitken, N., Lupart, J., & Scott, S. (2009). The Alberta student assessment study: Final report. Edmonton: Alberta Education. Retrieved Feb. 25, 2014 from <http://education.alberta.ca/media/1165612/albertaassessmentstudyfinalreport.pdf>

- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21(1), 1–19.
- Weinbaum, E. H. (2009). Learning about assessment: An evaluation of a ten-state effort to build assessment capacity in high schools. *Consortium for Policy Research in Education (CPRE)*. CPRE Research Report #RR-61. Retrieved Feb.16, 2014 from http://cpre.org/sites/default/files/researchreport/827_cpreten-stateassessmentweb-copy.pdf
- Weinmann, T., Thomas, S., Brilmayer, S., Heinrich, S., & Radon, K. (2012). Testing Skype as an interview method in epidemiologic research: response and feasibility. *International Journal of Public Health*, 57(6). Retrieved Dec. 12, 2012 from <http://dx.doi.org/10.1007/s00038-012-0404-7>
- Wenglinsky, H. (1997). How money matters: The effect of school district spending on academic achievement. *Sociology of Education*, 70(3), 221. Retrieved Apr. 13, 2013 from <http://dx.doi.org/10.2307/2673210>
- Wenglinsky, H. (2000). How teaching matters: Bringing the classroom back into discussions of teacher quality. Princeton, NY: Milken Family Foundation and Educational Testing Service. Retrieved Feb. 16, 2013 from <http://www.ets.org/Media/Research/pdf/PICTEAMAT.pdf>
- Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, 10(12), 1–30.
- Wideman, R. (1999). The power of action research: Supporting professional learning through inquiry. *The Ontario Action Researcher*, 2(2). Retrieved Mar. 30, 2013 from <http://oar.nipissingu.ca/archive-Vol2No2-V223E.htm>
- Wideman, R. (2003). Using action research and provincial test results to improve student learning. *International Electronic Journal for Leadership in Learning*, 6 (Jan. 2001).
- Wideman, R. (2011). Empowering teachers and schools to play their key role in improving education. *Canadian Journal of Action Research*, 12(3), 47–59.
- Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, 45(2), 107–122.

- Willms, J. D. (2000). Monitoring school performance for 'standards-based reform'. *Evaluation and Research in Education, 14*.
- Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*(1992), 327–350.
- Witt, J. C., Noell, G. H., LaFleur, L. H., & Mortenson, B. P. (1997). Teacher use of interventions in general education settings: Measurement and analysis of the independent variable. *Journal of Applied Behavior Analysis, 30*.
- Woessmann, L. (2001). Why students in some countries do better: International evidence on the importance of education policy. *Education Matters, 1*(2), 67–74. Retrieved Mar. 31, 2013 from http://educationnext.org/files/ednext20012_67.pdf
- Wößmann, L. (2003). Central exams as the “currency” of school systems: International evidence on the complementarity of school autonomy and central exams. *CESifo Group. CESifo DICE Report 4/2003* (Vol. 72, pp. 46–56). Retrieved Feb. 16, 2013 from <http://www.cesifo-group.de/portal/pls/portal/docs/1/1193688.PDF>
- Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Kieler Arbeitspapiere, No. 98*. Retrieved Sep. 16, 2014 from <http://www.econstor.eu/bitstream/10419/17917/1/kap983.pdf>
- Yeh, S. S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. *American Educational Research Journal, 30*(9).
- Yeh, S. S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives, 13*(43).
- Yeh, S. S. (2010). The cost effectiveness of NBPTS teacher certification. *Evaluation Review, 34*(3), 220–41. doi:10.1177/0193841X10369752
- Yen, W. M., & Henderson, D. L. (2002). Professional standards related to using large-scale state assessments in decisions for individual students. *Measurement and Evaluation in Counselling and Development, 35*, 132–143.
- Young, V. M. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education, 112*(4), 521-548.

- Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives*, 18(19), 1–40.
- Youngs, P. (2001). District and state policy influences on professional development and school capacity. *Educational Policy*, 15, 278–301.
- Zigarelli, A. (1996). An empirical test of conclusions from effective schools research. *The Journal of Educational Research*, 90(2), 103–110.
- Zigo, D. (2001). Constructing firebreaks against high-stakes testing. *English Education*, 33(3).

Appendix (1) Code of professional competence ⁹⁸

Reproduced in its entirety, this is the professional code of competence from the province of Saskatchewan. Similar documents are produced and implemented in other jurisdictions in Canada and internationally (Nuland & Poisson, 2009). The relevant sections for determining reactivity effects are in bold type below:

Code of Professional Competence

The Code of Professional Competence applies to members of the Saskatchewan Teachers' Federation. The code includes the following core principles of competent teaching practice, each of which teachers may demonstrate in various ways.

- To create and maintain a learning environment that encourages and supports the growth of the whole student.
- To demonstrate a professional level of knowledge about the curriculum and the skills and judgment required to apply this knowledge effectively.
- To demonstrate and support a repertoire of instructional strategies and methods that are applied in teaching activities.
- To carry out professional responsibilities for student assessment and evaluation.
- To reflect upon the goals and experience of professional practice, and adapt one's teaching accordingly.
- To work with colleagues in mutually supportive ways and develop effective professional relationships with members of the educational community.
- To conduct all professional relationships in ways that are consistent with principles of equity, fairness and respect for others.

⁹⁸ This document can be found online at:

<https://www.stf.sk.ca/portal.jsp?Sy3uQUUnbK9L2RmSZs02CjVy0w7ZkI/ks6g2u00gzAtsk=F#portal.jsp?S3ua0P4leiBvLe5BSdsr0vZGZJmzTYKNX8t/KNvKOzGyZacpsswpYUA==F>

Appendix (2) Teacher Survey

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.

Testing and teaching in my classroom – Nova Scotia

Page 1

When provincial governments enacted the large scale testing of students, teachers were designated to prepare students and to monitor the assessments. Teachers were also, according to policy documents, expected to pay attention to the results and to use data to improve their instruction. How these data are used in practice by the professional educators in Canada is a question that bears asking. What is clear is that front-line teachers have strong opinions based in their experience about the role of large scale testing in their classrooms.

I am a teacher and a PhD candidate studying large-scale provincial assessment in Canada. The best way for me to try to understand how teachers use these provincial test data is to ask them directly, and this survey does that - it asks you about your experience with provincial tests in your school. The survey will take approximately 20 minutes to complete, but less time if you do not give provincial tests.

For Nova Scotia teachers, the relevant provincial assessments would be the annual Nova Scotia Assessment tests in English, French and Mathematics for grades 3, 4, 6 and 8, as well as grade 10 tests in English, French, and Mathematics starting in 2013-2014. Divisional or school-initiated diagnostic tests (such as Fountas and Pinell) and national / international achievement tests (such as PCAP and PISA) are NOT the subject of this questionnaire.

There is complete confidentiality and anonymity for all respondents to this survey. The results will not be released for any purpose other than writing my dissertation. You can supply your email address if you are interested in the possibility of a follow-up interview, but your data will remain confidential.

* Survey Monkey is a web-survey company located in the USA, and is the host of this on-line research. This company is subject to the US Patriot Act that allows authorities access to the records of internet service providers. Survey Monkey's servers record incoming IP addresses – including that of the computer that you use to access the survey. However, no connection is made between your data and your computer's IP address. If you choose to participate in the survey, you understand that your responses to the survey questions will be stored and accessed in the USA.

** The researcher can be contacted via email with comments or concerns from the last page of the survey.

***1. By clicking 'yes' below I am consenting to the terms of this voluntary survey.**

Yes	No
<input checked="" type="radio"/>	<input type="radio"/>

Page 2

Biographical Information: Answers to the following questions are the only specific personal information in the survey.

Keep in mind that all biographical information will be kept private and confidential. Providing your email in the final question is a matter of your own personal choice and will not alter the fact that all information is confidential in nature.

***2. What is your age?**

<input checked="" type="radio"/> 18 to 24 [1]	<input type="radio"/> 25 to 34 [2]	<input type="radio"/> 35 to 44 [3]	<input type="radio"/> 45 to 54 [4]	<input type="radio"/> 55 to 64 [5]	<input type="radio"/> 65 or older [6]
-----------------------------------------------	------------------------------------	------------------------------------	------------------------------------	------------------------------------	---------------------------------------

***3. What is your gender?**

<input type="radio"/> Female [2]	<input type="radio"/> Male [1]
----------------------------------	--------------------------------

Testing and teaching in my classroom – Nova Scotia

***4. Primary grade(s) taught: (Note: these grade groupings may not align with provincial or division / district grade groupings)**

- Kindergarten or Pre-K [1]
- Elementary (grades 1-5) [2]
- Middle Years (grades 6-8) [3]
- High School (grades 9-12) [4]

***5. How many years have you been teaching? (Based on full time years of service.)**

- 0-4 years [1]
- 5-9 years [2]
- 10-14 years [3]
- 15-19 years [4]
- 20-24 years [5]
- 25 or more years [6]

***6. How would you characterize your school?**

- It is an urban school [1]
- It is a suburban school [1]
- It is a rural school [2]
- It is a northern or remote school [2]

7. How many teachers are on staff at your school?

- Less than 15 [1]
- 15 - 24 [2]
- 25 - 34 [3]
- 35 - 44 [4]
- 45 or more [5]

***8. How many students on average are in your class(es)?**

- 1-10 students [1]
- 11-15 students [2]
- 16-20 students [3]
- 21-25 students [4]
- 26-30 students [5]
- More than 30 students [6]

Testing and teaching in my classroom - Nova Scotia

*9. What is the level of your highest educational qualifications?

- College or high school [1]
- University degree without completed B.Ed. [2]
- University degree including B.Ed. [3]
- Graduate studies without completed Master's [4]
- Completed Master's degree or more [5]

*10. In which subject areas are your university credentials? (Please note that the choices are arranged in terms of those subjects which have provincial tests, not to delegate non-core subjects as less important.)

	Major	Minor
English	<input type="checkbox"/>	<input type="checkbox"/>
Mathematics	<input type="checkbox"/>	<input type="checkbox"/>
Science	<input type="checkbox"/>	<input type="checkbox"/>
Social Sciences	<input type="checkbox"/>	<input type="checkbox"/>
Other	<input type="checkbox"/>	<input type="checkbox"/>

*11. School Division / District: (For your current position.)

12. If you would not mind being contacted for more information in the future, please include your email address.

*13. Testing in Your Classroom:

- I give provincial achievement tests in my classroom.
- I do not give provincial achievement tests in my classroom.

Page 3

The Tests: The following section asks general questions about the tests, and how and when the tests are given.

Testing and teaching in my classroom – Nova Scotia

*** 14. Which provincially-mandated large-scale assessments does your class write? Include tests written annually or bi-annually that have been written within the last 2-3 years.**

	Core English or French	Mathematics	Science	Social Studies	Other
Kindergarten or Pre-K	<input type="checkbox"/>				
Elementary (grades 1-5)	<input type="checkbox"/>				
Middle Years (grades 6-8)	<input type="checkbox"/>				
High School (grades 9-12)	<input type="checkbox"/>				

*** 15. When are the test results returned to you?**

- During the same school year the test is written [1]
- The next school year [-0.5]
- The results are not returned to me [-1]
- I'm not sure [-1]

*** 16. Do the results you see from provincial tests compare:**

	Yes [1]	No [-0.5]	I'm not sure [-1]	I do not see these results [-1]
Your division / district with other divisions / districts or average scores	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Your school with other schools or average scores	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Teachers / classes with other teachers / classes or average scores	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Students with other students or average scores	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*** 17. In what ways are the results on provincial assessments given to staff members?**

- They are presented by department heads [1]
- They are presented by administration [1]
- They are presented by divisional / district personnel [1]
- They are presented by ministry personnel [1]
- I do not see the results for my class or the school [-1]
- I'm not sure [-1]
- Other (please specify)

Testing and teaching in my classroom – Nova Scotia

***18. Consider the manner in which assessment results are presented to you. Is the information reported 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 the results as presented. [-1]
- I do not see the results. [-1]

***19. Is it possible for you and other teachers in your community to use these assessment results directly to inform your instruction?**

- Yes, we can act on the results directly. [1]
- Some interpretation and analysis is needed before we can act. [-0.5]
- We cannot act directly because teachers are responsible for all the interpretation and analysis. [-1]
- We cannot act on the results because they are poorly or incompletely presented. [-1]
- Other (please specify)

***20. For each type of test item below, indicate whether you think that kind of questioning is used too often, not often enough, or the proper amount on provincial tests. Make one choice for each item type.**

	This type of item is used too much	This type of item is used too little	I think the current use is appropriate
Selected response (multiple choice, true/false, matching, fill ins with word bank)	<input type="radio"/> [-1]	<input type="radio"/> [-1]	<input type="radio"/> [1]
Short constructed response (fill ins without word bank, short answer, labeling, numerical response)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Longer constructed response (long answer, showing work, essays, performance tasks, etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Page 4

Using the data: This section asks about teachers' exposure to and use of provincial assessment data in their instruction.

***21. Have you ever been involved in the writing of questions or marking items for provincial assessments?**

- Yes
- No

Testing and teaching in my classroom – Nova Scotia

***22. Think about the ways your instruction may have changed in classes which write provincial assessments (as compared to classes that do not write these tests). Choose a response for all of the following statements:**

	[0] Not at all	[0.5] Somewhat	[1] A great deal
I have looked for PD to improve my instructional strategies	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have requested additional resources related to testing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have worked with other teachers to make sense the data	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
I cover a wider range of topics in the curriculum	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I hold group study sessions or provide extra help after school	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

***23. Once more, consider the ways your instruction may have changed in classes which write provincial assessments and choose a response for all of the following statements:**

	Not at all	Somewhat	A great deal
I cover material I know will be on the test very thoroughly	<input type="radio"/> [0]	<input type="radio"/> [-0.5]	<input type="radio"/> [-1]
I focus more on test taking strategies like the "process of elimination"	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I use the format of the test to give similar types of practice questions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I focus more on subjects that have provincial tests	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I review old exam questions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

***24. Below is a list of responsibilities that commonly fall upon teachers. Check beside each statement whether or not you think it is an appropriate use for provincial assessments:**

	[1] Appropriate	[-1] Not Appropriate
Assign or reassign students to classes or groups	<input type="radio"/>	<input type="radio"/>
Identify learning needs of students who are struggling	<input type="radio"/>	<input type="radio"/>
Discuss student progress or instructional strategies with other educators	<input type="radio"/>	<input type="radio"/>
Form small groups of students for targeted instruction	<input type="radio"/>	<input type="radio"/>
Discuss data with a parent	<input type="radio"/>	<input type="radio"/>
Discuss data with a student	<input type="radio"/>	<input type="radio"/>
Choose which parents to contact	<input type="radio"/>	<input type="radio"/>
Meet with a specialist about data - e.g., instructional coach	<input type="radio"/>	<input type="radio"/>

Testing and teaching in my classroom - Nova Scotia

***25. For the instructional strategies below that you use in class, indicate whether your use of that strategy differs from class to class. The 'performance' and 'results' below refer to both provincial tests and other classroom assessments:**

	Only in tested classes	Only in non-tested classes	In all classes
Plan different assignments based on test performance	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Re-teach topics when performance on assessments did not meet expectations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Review test results to identify students who need extra instruction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Review test results to identify topics needing more or less emphasis	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Confer with another teacher about alternative ways to present topics / lessons	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Page 5

Supports for Testing: These questions ask about the supports teachers have in using provincial assessment data, and also the supports they find most helpful (or would if they were provided).

***26. Do you share and discuss provincial test results with your students' teachers for the following year?**

- Never [-1]
- Sometimes [0]
- Always [1]

***27. Have results from previous years' provincial tests been shared and discussed with you?**

- Never [-1]
- Sometimes [0]
- Always [1]

Testing and teaching in my classroom – Nova Scotia

***28. What supports do you use to assist you with the assessment results? Check all that apply.**

	[1] Provided by my school	[1] Provided by my district / division	[1] Provided by my ministry
Professional Development	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Professional Learning Communities	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Assessment Teams	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Administrative support	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Printed or online guides	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Coaching and / or mentoring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Other (please specify)

***29. Considering all the supports you receive, who provides you the most help in your use of provincial assessment data? Check one option for each statement:**

	Very helpful [1]	Helpful [0.5]	Not helpful [-0.5]	Not provided [-1]
School supports are:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
District / division supports are:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ministry supports are:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Page 6

Incentives: This section refers to stakes or incentives applied to provincial tests and also to any expectations that test results are used to inform instruction.

***30. Teachers may be expected or pressured to use the assessment results to improve classroom instruction. Check any (or all) of the following statements that best matches your experience.**

- I am expected to use the results by school administration [1]
- I am expected to use the results by my division / district [1]
- I am expected to use the results by the ministry [1]
- There is no expectation that I use the results [-1]

Testing and teaching in my classroom – Nova Scotia

***31. Who follows up on your use of provincial assessment results to improve or change instruction? Check all that apply.**

- School administration [1]
 Division / district staff [1]
 Ministry staff [1]
 There is no follow up on instructional change or improvement [-1]
 Other (please specify)

***32. How much pressure do you feel to improve your students' results on these tests?**

- None [0]
 A small amount [0.5]
 A great deal [1]

***33. To the best of your recollection how were your results from the last provincial assessments?**

	[1] Better than average	[1] About the same as the average	[1] Worse than average	[-1] I don't recall	[-1] These results are not provided
My class results were:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My school results were:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My division / district results were:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

***34. Considering that provincial test results are used formally or informally to rank schools, agree or disagree with the following statement:**

	Agree	Neither agree nor disagree	Disagree
Provincial tests are an appropriate way to measure and compare classes, schools and divisions / districts.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

***35. I would qualify the provincial assessments in my jurisdiction as: (Note that this refers to incentives applied to teachers and / or schools - NOT students.)**

- High stakes tests for teachers and / or schools [1]
 Medium stakes tests for teachers and / or schools [0.5]
 Low stakes tests for teachers and / or schools [0]

Page 7

Purposes of Testing: The questions that follow relate to your general opinions and impressions of provincial testing.

Testing and teaching in my classroom – Nova Scotia

Please answer them as best you can based on your knowledge and experience.

*36. For School Accountability, provincial testing:

	[-1] Disagree	[0] Neither agree nor disagree	[1] Agree
Is a good way to evaluate a school	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is an accurate indicator of a school's quality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*37. For Student / Parent Accountability, provincial testing:

	[-1] Disagree	[0] Neither agree nor disagree	[1] Agree
Determines if students meet qualification standards	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Makes parents better aware of student growth	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Selects students for education / employment opportunities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*38. For School Improvement, provincial testing:

	[-1] Disagree	[0] Neither agree nor disagree	[1] Agree
Identifies student strengths and weaknesses	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Helps students improve their learning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is integrated with teaching practice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Allows different students to get different instruction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Changes the way teachers teach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*39. Provincial testing:

	[1] Disagree	[0] Neither agree nor disagree	[-1] Agree
Interferes with appropriate teaching	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Data only get used when stakes are high	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has little impact on teaching practices	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Results are ignored and filed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is an imprecise process	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Thank You!

Click 'done' below to record your responses. I greatly appreciate your time in completing this survey.

Questions or concerns can be directed to:
dcopp.gssd@gmail.com

Appendix (3) Interview Guide

This is the interview guide used by the researcher. Since the format of the interviews was semi-structured, many other unrecorded follow up questions were asked when circumstances were appropriate.

Introduction

Name:

Position:

Province:

Consent to record:

Test and Attitudes

Can you describe the provincial tests your students write?

What are the best features of this test?

What are the limiting factors of this test?

When do you see the results, and how are they presented to you?

Describe the results that you see for the assessment.

Are the results used to compare classes? Schools? Divisions?

Are provincial tests necessary for accountability in schools?

Do you have any concerns about this model of school accountability?

Supports

How common are discussions about test results in your school? Are they about how to improve scores or how to improve instruction?

What supports are provided in your work with these assessments? Which supports do you find most useful?

Does your school or division make time for you to work with these results?

What supports are not provided enough or would help but aren't available to you?

Incentives

* Describe how (or if) you are expected to use the results from provincial tests.

Who follows up on this expectation? How often?

What has been the trend in test scores for your class/school over the last 3-5 years?

To what factor(s) is this trend attributed?

Is there much pressure applied to teachers to improve provincial test scores?

Reactivity

Describe how, in your experience, teachers react to the release of test results in your school.

* How do you use the data to inform how or what you teach?

* Can you describe one example of a change in instruction you have made as a result of seeing provincial test performance?

* Are there inappropriate uses of results data that you have witnessed?

* Do assessments like these improve teaching at schools? (How?)

Appendix (4) Interview coding key

The codes for interviews were established for the most part by the content of the survey. The purpose of the interviews was to follow up on survey responses with teachers, administrators and division-level staff. To gain more insight, teacher respondents were purposefully selected for their positive and negative reactivity survey responses. Interview quotations were then evaluated (based on their probative value) for their fitness to be included in the related chapters.

#	IVs	#	DVs
1	time of data return	21	positive reactivity
2	aggregated/disaggregated data	22	negative reactivity
3	item types	23	reactivity effects
4	presentation of results	#	BFVs
5	data clarity	31	age
6	ability to act on data	32	sex
7	results use for school accountability	33	grade taught
8	results use for student accountability	34	experience
9	results use for school improvement	35	setting (urban or rural)
10	negative attitudes about testing	36	staff size
11	appropriate uses for data	37	class size
12	sharing of data	38	qualifications
13	school supports	#	Others
14	division supports	41	teacher autonomy
15	ministry supports	42	relationships with students
16	expectations	43	comparing/rating schools
17	follow up	44	comparing/rating teachers
18	results awareness	45	costs / wasted resources
19	perceived pressure		
20	perceived stakes		

MSGoG Dissertation Series

Michaella Vanore

Family-Member Migration and the Psychosocial Health Outcomes of Children in Moldova and Georgia

MGSOG Dissertation Series, nr 57 (2015)

Sonja Fransen

The Economic and Social Effects of Remittances and Return Migration in Conflict-Affected Areas: The Case of Burundi

MGSOG Dissertation Series, nr 56 (2015)

Ibrahim Khalil Conteh

Natural Hazards and Education

The Impact of Floods on Primary School Education in Zambia

MGSOG Dissertation Series, nr 55 (2015)

Richard Bluhm

Growth Dynamics and Development

Essays in Applied Econometrics and Political Economy

MGSOG Dissertation Series, nr 54 (2015)

Nevena P. Zhelyazkova

Work-Family Reconciliation and Use of Parental Leave in Luxembourg

Empirical Analysis of Administrative Records

MGSOG Dissertation Series, nr 53 (2015)

Sachin Kumar Badkas

Metachoice and Metadata

Innovating with Environmental Policy Analysis in Europe

MGSOG Dissertation Series, nr 52 (2014)

Irina S. Burlacu

*An Evaluation of Tax-Benefit Systems Impact on the Welfare of Frontier Workers
The Case of Luxembourg and Belgium*

MGSOG Dissertation Series, nr 51 (2014)

Özge Bilgili

*Simultaneity in Transnational Migration Research:
Links Between Migrants' Host and Home Country Orientation*

MGSOG Dissertation Series, nr 50 (2014)

Yulia Privalova Krieger

*Reshaping the Big Agenda: Transnational Politics and Domestic Resistance
Financial crisis and social protection reform in Bosnia and Herzegovina*

MGSOG Dissertation Series, nr 49 (2014)

Marieke van Houte

*Moving Back or Moving Forward?
Return migration after conflict*

MGSOG Dissertation Series, nr 48 (2014)

Oxana Slobozhan

*Global Governance in the Management of Natural Resources
The Case of the Extractive Industries Transparency Initiative (EITI)*

MGSOG Dissertation Series, nr 47 (2014)

Luis Bernardo Mejia Guinand

*The Changing Role of the Central Planning Offices in Latin America:
A Comparative Historical Analysis Perspective (1950-2013)*

MGSOG Dissertation Series, nr 46 (2014)

Cheng Boon Ong

Ethnic Segregation in Housing, Schools and Neighbourhoods in the Netherlands

MGSOG Dissertation Series, nr 45 (2014)

Luciana V. Cingolani

*Bureaucracies for Development: Oxymoron or Reality?
Studies on State Capacity in Challenging Governance Contexts*

MGSOG Dissertation Series, nr 44 (2014)

Carlos Cadena Gaitán
Green Politics in Latin American Cities - Sustainable Transport Agendas
MGSoG Dissertation Series, nr 43 (2014)

Katie Kuschminder
Female Return Migration and Reintegration Strategies in Ethiopia
MGSoG Dissertation Series, nr 42 (2014)

Metka Hercog
Highly-Skilled Migration and New Destination Countries
MGSoG Dissertation Series, nr 41 (2014)

Margaret Agaba Rugadya
Can Remittances Influence the Tenure and Quality of Housing in Uganda?
MGSoG Dissertation Series, nr 40 (2014)

Ilire Agimi
*New Governance Under Limited Statehood
The Case of Local Government Reform in Kosovo*
MGSoG Dissertation Series, nr 39 (2014)

Kristine Farla
Empirical Studies on Institutions, Policies and Economic Development
MGSoG Dissertation Series, nr 38 (2013)

Marina Petrovic
*Social Assistance and Activation in the Pursuit of Happiness:
Shedding New Light on Old Policy Solutions to Social Exclusion*
MGSoG Dissertation Series, nr 37 (2013)

Laura Torvinen
*Assessing Governance Assessments; The Case of Mozambique
Governance Assessments in the Context of Aid Effectiveness Discourse*
MGSoG Dissertation Series, nr 36 (2013)

Biniam Egu Bedasso
*Institutional Change in the Long Shadow of Elites
Essays on Institutions, Human Capital and Ethnicity in Developing Countries*
MGSoG Dissertation Series, nr 35 (2013)

Sepideh Yousefzadeh Faal Deghati

Childhoods Embargoed

Constructing and Reconstructing Multidimensional Child Poverty in Iran 1984-2009

MGSOG Dissertation Series, nr 34 (2013)

Robert Bauchmüller

Investing in Early Childhood Care and Education:

The Impact of Quality on Inequality

MGSOG Dissertation Series, nr 33 (2013)

Martin Rehm

Unified Yet Separated

Empirical Study on the Impact of Hierarchical Positions within Communities of Learning

MGSOG Dissertation Series, nr 32 (2013)

Dorcias Mbuvi

Utility Reforms and Performance of the Urban Water Sector in Africa

MGSOG Dissertation Series, nr 31 (2012)

Lina Salanauskaite

Distributional Impacts of Public Policies:

Essays in Ex-Ante and Ex-Post Evaluation

MGSOG Dissertation Series, nr 30 (2012)

Esther Schüring

To Condition or not – is that the Question?

An Analysis of the Effectiveness of Ex-Ante and Ex-Post Conditionality in Social Cash Transfer Programs

MGSOG Dissertation Series, nr 29 (2012)

Joe Abah

Strong Organisations in Weak States

Atypical Public Sector Performance in Dysfunctional Environments

MGSOG Dissertation Series, nr 28 (2012)

Zina Samih Nimeh

Social Citizenship Rights: Inequality and Exclusion

MGSOG Dissertation Series, nr 27 (2012)

Lenka Eisenhamerová
Legitimacy of 'Humanitarian Military Intervention'
MGSoG Dissertation Series, nr 26 (2011)

Sonila Tomini
Informal Payments for Health Care Services in Albania
MGSoG Dissertation Series, nr 25 (2011)

Jinjing Li
Dynamic Microsimulation in Public Policy Evaluation
MGSoG Dissertation Series, nr 24 (2011)

Aziz Atamanov
Rural Nonfarm Employment and International Migration as Alternatives to Agricultural Employment: The Case of Kyrgyzstan
MGSoG Dissertation Series, nr 23 (2011)

Frieda Vandeninden
Poverty Alleviation: Aid and Social Pensions
MGSoG Dissertation Series, nr 22 (2011)

Juliana Nyasha Tirivayi
The Welfare Effects of Integrating AIDS Treatment with Food Transfers: Evidence from Zambia
MGSoG Dissertation Series, nr 21 (2011)

Agnieska Ewa Sowa
Who's Left Behind? Social Dimensions of Health Transition and Utilization of Medical Care in Poland
MGSoG Dissertation Series, nr 20 (2011)

Emmanaouil Sfakianakis
The Role of Private Actors in the Provision of Public Goods with Applications to Infrastructure and Financial Stability
MGSoG Dissertation Series, nr 19 (2011)

Siu Hing Lo

White Collars Green Sleeves

An Interorganizational Comparison of Determinants of Energy-Related Behaviors among Office Workers

MGSOG Dissertation Series, nr 18 (2011)

Treena Wu

Constraints to Human Capital Investment in Developing Countries:

Using the Asian Financial Crisis in Indonesia as a Natural Experiment

MGSOG Dissertation Series, nr 17 (2011)

Henry Espinoza Peña

Impact Evaluation of a Job-Training Programme for Disadvantaged Youths:

The Case of Projoven

MGSOG Dissertation Series, nr 16 (2011)

Florian Tomini

Between Family and Friends

Understanding the Interdependency of Private Transfers

MGSOG Dissertation Series, nr 15 (2010)

Michał Polalowski

The Institutional Transformation of Social Policy in East Central Europe:

Poland and Hungary in comparative and historical perspective

MGSOG Dissertation Series, nr 14 (2010)

Maha Ahmed

Defining, Measuring and Addressing Vulnerability:

The Case of Post Conflict Environments

MGSOG Dissertation Series, nr 13 (2010)

Pascal Beckers

Local Space and Economic Success

The role of spatial segregation of migrants in the Netherlands

MGSOG Dissertation Series, nr 12 (2010)

Victor Cebotari

Conflicting Demands in Ethnically Diverse Societies

Ethnopolitical Contention and Identity Values in Europe

MGSOG Dissertation Series, nr 11 (2010)

Dennis Gyllensporre

Competing and Complementary Perspectives on the EU as a Crisis Management Actor: An Examination of the Common Security and Defence Policy through the Lenses of Idealism and Realism

MGSOG Dissertation Series, nr 10 (2010)

Judit Vall Castello

Business Cycle and Policy Effects on Labour Market Transitions of Older and Disabled Workers in Spain

MGSOG Dissertation Series, nr. 9 (2010)

Keetie Roelen

False Positives or Hidden Dimensions: the definition and measurement of child poverty

MGSOG Dissertation Series, nr. 8 (2010)

Denisa Maria Sologon

Earning Dynamics in Europe

MGSOG Dissertation Series, nr. 7 (2010)

Melissa Siegel

Money and Mobility: Migration and Remittances

MGSOG Dissertation Series, nr. 6 (2010)

Jessica S. Hagen-Zanker

Modest Expectations: Causes and effects of migration on migrant households in source countries

MGSOG Dissertation Series, nr. 5 (2010)

Mirtha R. Muniz Castillo

Human Development and Autonomy in Project Aid: Experiences from four bilateral projects in Nicaragua and El Salvador

MGSOG Dissertation Series, nr. 4 (2009)

Christiane Arndt

Governance Indicators

MGSOG Dissertation Series, nr. 3 (2009)

Britta Augsburg

Microfinance – Greater Good or Lesser Evil?

MGSOG Dissertation Series, nr. 2 (2009)

Geranda Notten
Measuring and Managing Poverty Risks
MGSoG Dissertation Series, nr. 1 (2008)