The Canadian Journal of Program Evaluation Vol. 22 No. 3 Pages 73–97
ISSN 0834-1516

SECONDARY ANALYSIS WITH MINORITY GROUP DATA: A RESEARCH TEAM'S ACCOUNT OF THE CHALLENGES

Marielle Simon
Nicole Roberts
Robin Tierney
Renée Forgette-Giroux
University of Ottawa
Ottawa, Ontario

Abstract: Understanding the challenges associated with conducting secondary analysis of large-scale assessment data is important for identifying the strengths and weaknesses of various statistical models, and it can lead to the improvement of this type of research. The challenges encountered in the analysis of assessment data from subpopulations may be of particular value for this purpose. To date, few studies have discussed the problems associated with the secondary analysis of large-scale assessment data. By relating the experiences of a research team that engaged in several projects involving secondary analyses of linguistic minority population data from three different large-scale assessment programs, this article aims to help readers understand the practical, conceptual, and technical/statistical challenges that can be encountered.

Résumé : Il est important de comprendre les défis inhérents à l'analyse secondaire des données issues d'évaluation à grande échelle du rendement d'élèves afin d'identifier les forces et faiblesses de divers modèles statistiques ainsi que d'améliorer ce genre de recherche. À ce jour, peu d'études se sont penchées sur les problèmes associés à l'analyse secondaire des études à grande échelle. L'analyse des résultats de sous-populations pose des problèmes particuliers. En relatant les expériences d'un groupe de chercheurs engagés dans plusieurs projets utilisant des résultats d'analyses de données secondaires de populations linguistiques minoritaires provenant de trois programmes d'évaluation à grande échelle, cet article vise à aider les lecteurs à comprendre les défis pratiques, conceptuels, techniques, et statistiques qu'ils peuvent rencontrer.

Corresponding author: Marielle Simon, Faculty of Education, University of Ottawa, 145 Jean-Jacques Lussier, P.O.Box 450 Station A, Ottawa, ON, K1N 6N5; <msimon@uottawa.ca>

INTRODUCTION

Large-scale assessment programs are typically viewed as important means of determining student achievement in educational systems, and they are normally carried out in a systematic fashion. While such assessments are a complex undertaking, they provide crucial data for making decisions regarding education. However, the wealth of data obtained from large-scale assessments is currently not being used to its fullest capacity in directing educational policy, research, or practice in Canada (Crocker, 2002; Levin, 2003). Calls for the secondary analysis of quantitative data have come from a number of organizations, including the Social Sciences and Humanities Research Council (SSHRC) and the Canadian Education Statistics Council (CESC), as well as from provincial, national, and international testing agencies (Gorard, 2002). Secondary data analysis is defined here as a set of research endeavours that use existing data to answer research questions that may or may not have been proposed when the data were originally collected (Rew, Koniak-Griffin, Lewis, Miles, & O'Sullivan, 2000).

Numerous advantages in conducting secondary data analysis have been documented in the literature. Large-scale testing agencies have suggested that secondary data analysis can be used to link contextual factors, educational factors, and student performance on provincial assessments (Rogers, Anderson, Klinger, & Dawber, 2006). Rew et al. (2000) note that the main advantage to secondary analysis is economical because data collection is usually the most time-consuming and expensive component of research, and the use of existing data saves time and money. Although access to some data sets involves a large financial fee, many are available free of charge (Gorard, 2002). Hofferth (2005) also suggests that secondary analysis is beneficial because it usually involves large sample sizes. It has been argued that more precision in answering research questions is possible with large data sets because they permit the use of advanced statistical techniques, such as Hierarchical Linear Modeling (HLM), that are not possible with smaller samples (Rogers et al., 2006).

Although the potential of secondary data analysis to meaningfully inform educational policy is stressed in the literature, there are a number of issues that are not immediately apparent. Reports of large-scale educational assessments in the news suggest that the effectiveness of teachers and schools can be studied using statistics, and this approach is often assumed by the general public to produce accurate

and objective results. However, it is also well known that statistics can be misused, misapplied, and misinterpreted. Stakeholders who use the results of large-scale assessments need to look more closely at the limitations of the data to which sophisticated statistical techniques have been applied, as the origin of policy disputes has repeatedly been traced to questions about the underlying data and related analyses (Maier, 1999). Secondary data analysts have raised serious questions about the integrity of existing data and the subsequent validity and usefulness of the results and findings (Brooks-Gunn, Phelps, & Elder, 1991; Goldstein, 2004; Rogers et al., 2006).

Understanding the benefits and challenges associated with conducting secondary analysis of large-scale assessment data is important to ensure the validity of interpretations and to develop possible solutions that are both relevant and practical for more efficient use of existing data. This process is complicated by the multidisciplinary nature of the available information. For example, articles that focus on the process and issues of conducting secondary data analysis are situated in the areas of psychology (Brooks-Gunn et al., 1991; McCall & Appelbaum, 1991), nursing (Rew et al., 2000), family studies (Hofferth, 2005), and education (Goldstein, 2004; Gorard, 2002; Wang, 2001). The application of secondary analyses to minority group data sets poses even further complications, which occur at all stages of the analyses from accessing the data to reporting results.

This article builds on the work of Rogers and colleagues (2006), who outline six issues relating to technique and the nature of secondary analysis outcomes, which need to be recognized in order for secondary analysis to successfully inform practice in education. Rogers and colleagues initially reported on the pitfalls and potential of secondary analyses with data from the pan-Canadian School Achievement Indicator Program (SAIP) in a session at the Canadian Society for the Study of Education conference in 2004. Some of the members of the Measurement, Assessment, and Evaluation (MEA) research unit, from the University of Ottawa, who were also engaged in a few research projects involving secondary analyses, consequently decided to document their process of data analyses. Whereas Rogers and colleagues (2006) focused their commentary on data from the SAIP mathematics assessment, this article gives a retrospective account of problems that arose when secondary analysis was conducted with minority linguistic group data sets. This study reports the experiences of a bilingual research team in Canada as they analyzed existing data from three large-scale international and national assessment

programs. In particular, it illustrates real-life challenges faced by the researchers and their assistants, and suggests possible solutions to the obstacles that were experienced. Essentially, this article gives an account of a research team's journey into the world of secondary data analyses. The challenges described not only inform novice researchers in the field, but also help identify strategies that need to be considered for future large-scale assessment programs in order to maximize their utility for minority groups. The following sections provide a description of the broader context of the work, details on the original studies that precipitated this article, and a framework for the account. The challenges that were encountered in the original studies are organized and presented as three types: practical, conceptual, and technical/statistical. A narrative of each challenge encountered is followed by a retrospective response that draws on the literature relating to secondary data analysis.

Description of the Original Studies and Their Context

All provincial jurisdictions in Canada currently conduct at least one large-scale assessment of student achievement (Klinger, 2007). Many of these provinces also participate in international studies such as the Third International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Program for International Student Achievement (PISA). In addition to provincial and international assessments conducted in the various Canadian jurisdictions, the Council of Ministers of Education, Canada (CMEC) provides leadership in education at the pan-Canadian and international levels and contributes to the fulfillment of the constitutional responsibility for education conferred on provinces and territories. According to CMEC, Canadians have long been interested in how well their education systems are meeting the needs of students and society. To provide information on these issues, the provinces and territories, through CMEC, developed the SAIP, which has since been replaced in 2005 by the Pan-Canadian Assessment Program (PCAP), to assess the performance of students in mathematics, reading, and science. Within most jurisdictions, students from both official linguistic communities are tested in their first language, English or French.

The principal researchers, Dr. Marielle Simon and Dr. Renée Forgette-Giroux, two of the founders of the MEA research unit, obtained two grants from the SSHRC, one in 2003 and another in 2004. The first study proposed to link variables from the TIMSS-Repeat (TIMSS-R) teacher background questionnaire with achievement of students

enrolled in French-language schools in Ontario. This French-language subpopulation represents approximately 5% of the provincial student population. The Canadian Charter of Rights ensures that the French- or English-language minority group in each province be allowed to receive its education in its language. Four French-language minority groups in Canada are large enough for their assessment results to be reported: Ontario, New Brunswick, Nova Scotia, and Manitoba. Conversely, Quebec has a large English-language minority group.

The second study was submitted a year later and proposed a comparative study of the links between teaching practices and student achievement on the PIRLS 2001 reading and SAIP 2002 writing III assessments from subgroup populations: Ontario French, Ontario English, and Quebec French. The underlying long-term motivation for both studies was to investigate possible explanations for the consistent and significantly lower achievement results of students from the minority French-language community in Ontario compared to the majority population.

Framework for the Research Account

As the research team began to encounter challenges in the first of the two original studies, a literature search was conducted to determine whether problems with secondary data analysis of large-scale assessment data, especially minority group data, had been reported in previous studies. McCall and Appelbaum (1991) discuss the different steps involved in conducting secondary analysis of large, multivariate, longitudinal, interdisciplinary databases, whereas Brooks-Gunn et al. (1991) stress the opportunities, possible limitations, and application of the analysis of longitudinal national data and secondary data analysis of long-term developmental studies. Clarke and Cossette (2000) grouped problems found in the secondary analysis of data collected in nursing research into three categories: theoretical, practical, and methodological. They identified some methodological issues that are similar to those noted by Rogers et al. (2006), but some of the theoretical and practical issues identified by Clarke and Cossette were also relevant to the research team. Although these articles do not involve analysis of minority group data from educational assessments, they did give members of the research team further insight into the methodological issues they were facing.

More specifically in the field of education, Goldstein (2004) raises some methodological concerns about the conduct, analysis, and interpreta-

tion of results from the PISA study. Goldstein also discusses certain features of the results that raise questions about the adequacy of the data, stresses the failure to introduce a longitudinal component, and makes suggestions for improvement. Robitaille and Beaton (2002) compiled a volume in order to make available the findings from a number of secondary analyses that researchers in many of the TIMSS countries have carried out since the data were collected in 1995. While the topics covered in that volume illustrate the range of investigations that this kind of data makes possible, it is based solely on the experience of TIMSS researchers.

Since 2000 a handful of Canadian studies have conducted secondary analysis specifically on linguistic minority group data sets resulting from large-scale assessments. For example, Bouchamma and Lapointe (2007), Lévesque and Bouchamma (2006), and Savard, Sévigny, and Beaudoin (2007) applied secondary analyses on pooled SAIP writing data of students enrolled in French-language schools across Canada. Landry and Allard (2002) compared four subpopulations (Ontario French, Ontario English, Quebec French, and Quebec English) and linked student achievement results with teacher and student variables from each of the SAIP's language arts, mathematics, and science studies. Herry (2000) and Plouffe, Simon, and Loye (2005) were particularly interested in examining relationships among contextual variables and student achievement in TIMSS mathematics. Lastly, Simon, Turcotte, Ferne, and Forgette-Giroux (2007) reported secondary analyses of teaching practices from Ontario French, Ontario English, and Quebec French teachers using PIRLS background questionnaire data. Although this field is growing, most of the work thus far has concentrated on reporting results, and few of the researchers give details on the challenges encountered in the process of the analyses.

Narratives about the research process in a variety of areas can readily be found. Recent examples published in educational journals are narratives about doing research relating to women and post-secondary education (Kimpson, 2005), and to activism and environmental education (Malone, 2006). An early account of a collaborative interdisciplinary research project by Crow, Levine, and Nager (1992) provided rough guidance for structuring this paper. Narratives about research that involves quantitative data are still not evident, though, and few tales have been told about conducting secondary data analysis. Data sources for this account include minutes generated at biweekly team meetings, research notes, printouts of e-mail dialogue between research assistants and researchers, and the ongoing reflections of

the research team members. In addition to the categories suggested by Clarke and Cossette (2000) and Rogers et al. (2006), the success and hindrance factors listed on the SSHRC final research report form that were also considered as challenges were identified and analyzed. Throughout the study, texts on secondary analysis (e.g., Kiecolt & Nathan, 1985) and other works involving secondary analysis of large-scale assessment data (e.g., Robitaille & Beaton, 2002) were consulted to aid in determining how to resolve problems and move forward.

In addition to the two principal researchers, many students were involved in conducting the two studies. Eight students participated, at one time or another, in the development of the conceptual frameworks, in the retrieval of the data, in analyzing the data, and in reporting results. Most contributed in their own way to the overall documentation of the issues faced in conducting secondary analyses on minority population data. The "we" in the narrative account refers to the two principal researchers and the students involved in the studies, two of whom are listed as authors of this article. The narrative account is provided in italics, followed by a discussion based on the relevant literature.

THE CHALLENGES ENCOUNTERED

This section provides a narrative account of the practical, conceptual, and technical/statistical challenges we encountered in conducting secondary analyses on three data sets over the period of four years.

Practical Challenges

We uncovered four practical challenges in conducting secondary analyses on minority group data sets. These issues were related to resources, data access, student training, and dissemination.

a) Resources

In 2003, the SSHRC in collaboration with Statistics Canada launched a joint funding program called the Initiative of the New Economy. It granted funding for approximately 10 studies per year for two years in a row. These studies all shared the secondary analysis of existing large-scale data sets available from Statistics Canada or the CMEC. Given our desire to investigate reasons for the significantly lower achievement results of students enrolled in Ontario French-language schools, we decided to submit a proposal. Our proposed study was

ambitious, attempting to compare secondary analyses of achievement results of minority groups from four provinces with those of the respective majority groups on the TIMSS, PISA, PIRLS, and SAIP data. We received $40,000, which was half of the requested grant, with no specific recommendations except to conduct pilot testing of secondary analyses. It was therefore up to us, the research team, to decide how to revise the study in light of this funding. Restricted funding and other difficulties, described later, shaped our decision to narrow the study to investigate TIMSS math data with only the Ontario French-language subpopulation.

Similar to the above experience, Beaton and Robitaille (2002) further discuss the difficulty that researchers face when attempting to allocate funding for such endeavours. Arguably, it seems easier to obtain funding to collect new data than to carry out more in-depth analysis of data already collected. Unless funding is made available for such in-depth secondary analysis of data from previous studies, the current state of affairs is unlikely to change and researchers will not be encouraged to take advantage of the many opportunities to analyze numerous large data sets already available or use their considerable potential value to address new questions and expand the breadth of existing knowledge. If funding is granted, it usually consists of small-scale, short-term funding. Short-term funding limits stability of research environments and offers researchers few occasions to develop networking. Moreover, researchers may be discouraged from conducting secondary data analysis, even after witnessing the advantages, because of the specific resources required, such as large financial investments for computer hardware, software, and training (Beaton & Robitaille, 2002).

b) Data access

International large-scale assessment programs normally provide Internet access to the data. However, this is not necessarily the case with subgroup data. In fact it was a major obstacle in the case of our TIMSS math study. According to our records, we were awarded our research funding on March 1, 2003. Subsequently, in an attempt to retrieve data codes, our e-mail and telephone messages were repeatedly sent back and forth to the Education Quality and Accountability office, to the TIMSS-R headquarters in British Columbia, and to Education Testing Services in New Jersey. Finally, on June 14, 2004, 15 months later, we received an e-mail from a Statistics Canada officer stationed in Germany that finally provided the Ontario French-language school codes.

Accessing the PISA data for the second study also proved tedious. Given the sensitive nature of the data and the relatively small sample sizes, we were told that data were only available in centres designated and run by Statistics Canada. Despite the fact that we were situated in Ottawa, where the main Statistics Canada offices are located, we could only access the data in a designated Montreal location.

Researchers should be aware that accessing secondary data may not be straightforward, particularly when dealing with minority groups. Another presenting issue is that the principal investigator may not have full control over all aspects of design and measurement because it is government- or organization-funded (Hofferth, 2005). Once data sets are retrieved, researchers often have difficulty obtaining answers to their specific questions. Is it possible to know how special cases were treated within the data? It is often difficult to locate such answers, and unknown vital information, such as whether children with learning disabilities were included in the sample, could compromise a researcher's hypothesis and analysis (McCall & Appelbaum, 1991). For security purposes, data related to subpopulations are not always as readily accessible as those associated with the dominant population.

c) Student training

Once we received funding, it was time to hire research assistants. For the two studies, we granted the equivalent of 12 contracts for 160 hours per session to eight students: five admitted in the measurement and evaluation concentration (three masters, one doctoral, and one student with special status) and three non-measurement students (two masters student, one studying in second-language and the other from educational counselling, and a doctoral student enrolled in teaching and learning). Two spoke mainly French, four spoke mainly English, and two were fully bilingual. Four had taken courses in statistics but had little experience in applying advanced statistical models to large-scale data sets. Only the student in the teaching and learning concentration accepted contracts as a full-time student throughout the entire period in which the studies were conducted (from 2002 to 2006). The others did a maximum of three assistantships consecutively.

Despite reliable access to students to serve as research assistants, this scenario presented specific challenges. Our research assistants' level of engagement in the study varied depending on timing and status in the program. For instance, some students were near program completion and others were starting. While some were full-time, others held part-

time positions. Such situations affected the investment the principal researchers were prepared to make in training the students. The lack of students specializing in measurement and advanced statistical skills represented a critical technical challenge. In the linguistic minority group context, the language barrier was yet another limitation that was difficult to resolve. At times, it meant having to resort to major interpretation and translation activities that were not initially planned in the research proposal.

The literature stresses that there is a need for more training opportunities in secondary analysis techniques in Canada, and perhaps such training sessions could be provided by the SSHRC or Statistics Canada (Beaton & Robitaille, 2002). For instance, Beaton and Robitaille suggest that training could be similar to the U.S. model of the National Center for Education Statistics (NCES) training advertised through the National Council for Measurement in Education (NCME). Such training seems necessary because few educational researchers have the methodological or statistical skills to handle both quantitative and qualitative data. Moreover, new data sets are often unfamiliar, and months of preliminary work are needed to become acquainted with the data (McCall & Appelbaum, 1991). There is a need for a mix of knowledge and skills related to large-scale studies, advanced statistics, and student assessment in order to successfully conduct secondary data analysis. Without an emphasis on training, researchers may be discouraged from conducting secondary analysis. Moreover, there will be many national variations and serious problems with comparability amongst large data sets. Training would ensure preparation of the next generation of experts in the field.

d) Disseminating results

The joint SSHRC and Statistics Canada granting program invited all grantees to a series of four symposiums. The first was offered in Ottawa in May 2003, the second was held in Quebec City in May 2004, the third took place in Ottawa in May 2005, and Victoria hosted the fourth and last symposium in 2006. The principal researchers, the co-researchers and up to four students participated in the Quebec 2004 and Ottawa 2005 symposiums. The two-day symposiums were closed to the general public and provided the grantees the opportunity to network among researchers and practitioners.

The audience consisted of decision-making officials from ministries of education across various provinces and territories, experts in measurement and evaluation, and officials from the granting agencies. This

mixed group resulted in two related challenges for us when it was time to present. First, it seemed that specific statistical and technical issues could not be properly addressed among researchers in this forum because of the mixed audience. Second, it was difficult to decide on the content and format of the presentations. Some of the questions we personally struggled with included "What level of statistics should be provided?", "What information is needed to interest policy makers?", "How should the presentation be worded to reach the policy makers, or the practitioners, or even the granting agency officials?", and "How can we balance the need to ensure statistical grounding and the need to provide meaningful information to ministry and granting agency officials?"

The discussion around the second issue persisted throughout the process of publication in referred journals. Additional questions relevant to the linguistic minority context involved whether to submit papers to French- or English-language journals. Unfortunately, there are few relevant French-language journals, and these mainly address the general public. The issue of language of publication resulted in having to translate various bits of work conducted by the French and English research assistants.

The literature on secondary analysis of large-scale assessment data does not address the issue of dissemination and publication. Perhaps our experience represents an isolated case, but it does highlight realities when dealing with bilingual contexts. When reporting results relevant to minority-language populations in their own language, the numbers of journals are limited. Publication issues also include the need to balance narrative and statistical reporting. However, forums for discussing specific issues around secondary analysis of large-scale assessment data sets, such as workshops and conferences, are gradually being supported by granting agencies, such as SSHRC, which is encouraging for those of us in the field.

Conceptual Challenges

This section covers the following three issues: conflicting study goals, sampling designs, and construct comparisons.

a) Goals

Each of our two studies initially stated specific research questions in their respective proposals. The TIMSS study asked the following three questions:

1. *What student characteristics influence Ontario French-language students' achievement on the TIMSS-R 1999 mathematics study?*
2. *What teaching and assessment practices impact on their achievement?*
3. *What is the relationship among the student characteristics, the teaching and assessment practices, and their impact on their achievement?*

The PIRLS/SAIP study attempted to answer the following research questions:

1. *What teaching and assessment classroom practices taken from the background questionnaires of the PIRLS 2001 and the SAIP 2002 studies seem to have a significant link with the reading and writing performance of Anglophone and Francophone students in Ontario and of Francophone students in Quebec?*
2. *What is the nature, diversity, and frequency of use of these practices?*
3. *To what extent do these teaching/assessment practices influence student performance in reading and writing, regardless of the targeted population?*
4. *How do these links differ amongst the three populations studied?*

The study of the background questionnaires proved disappointing when it came to obtaining data on teachers' assessment practices. The TIMSS math background questionnaire contained only 2 questions out of 20 related to assessment practices. Specifically, question 19 asked about the importance attached to seven assessment types such as standardized tests and student responses in class. Item 20 asked teachers to indicate how often they used assessment data for six different purposes, such as to calculate final grades and to plan their next lessons. Two other questions only briefly referred to frequency or time spent on assessment-related activities, such as the use of textbooks to find assessment items.

Pertaining to the PIRLS questionnaire, 3 questions out of 44, specifically items 26 to 28, were combined in a section titled "Assessment." These essentially asked about emphasis placed on various sources to monitor students' progress in reading, such as diagnostic tests, about frequency of use of different types of assessment such as multiple-choice

questions and about the specific use of portfolio in terms of assessment. No other questions referred to assessment.

SAIP writing III referred to assessment on 3 questions out of 33. Question 25 asked about weight given to 14 activities such as standardized tests produced outside the school, effort, and peer evaluation, and Question 26 referred to the number of different scores or grades used in computing final marks for English Language Arts students. Finally, item 13 asked to what extent teachers agreed with the statement "Assessment must be an integral and ongoing part of the learning process, not limited to final products." Five other questions throughout the test referred briefly to assessment but only in terms of time or frequency of use of assessment. Faced with this situation, we had to expand our definition of assessment to composites of items regarding various forms of questioning and homework.

One of the major limitations documented in the literature facing researchers doing secondary data analyses is related to the fit between the research question and the data set (Hofferth, 2005). Often, surveys may not contain precise indicators of the concepts the secondary analyst wants to study (Kiecolt & Nathan, 1985; Rew et al., 2000). The secondary data analysts are thus disadvantaged because they have to ensure the appropriateness of items and scales to their research questions and work closely with the data in order to create their hypotheses rather than use the ideal process that begins with a grounded research question (Hofferth, 2005). Researchers and policy makers or assessment officials should both be involved and consulted throughout the program from its conception to reporting results (Rogers et al., 2006).

b) Sampling designs

Our two secondary analysis studies both dealt with minority group data. Both looked at responses from students and their teachers. PISA does not gather teacher data. TIMSS-R, PIRLS, and SAIP do but TIMSS-R and PIRLS sample whole classes of students whereas the SAIP samples individual students based on age. Given that the original sampling was at the student level, secondary analysis on teacher data and resulting generalization of results were critically constrained.

Unsurprisingly, the kind of research that can be done with data is highly dependent upon the sampling design employed in the study,

and that design reflects developments in international assessments and assessment in general over many years (Beaton & Robitaille, 2002). Some researchers suggest that the potential for secondary data analysis would be enhanced if the lowest sampling unit was changed from student to class. Selection of all the classes in a school would enable further analysis; thus samples in secondary data analysis would be representative of their respective populations and of sufficient size to enable estimates that could be validly interpreted and generalizable (Rogers et al., 2006). However, in the case of minority populations, which are already oversampled, this suggestion remains problematic.

Another sampling issue relates to how some international studies of student achievement use age cohorts, and others use grade cohorts. Which is the more useful approach remains a debate, with advantages and disadvantages to each (Porter & Gamoran, 2002). Age-based samples make studying education effect more difficult because students are spread across a number of grades, making curriculum-based achievement tests problematic. Age-based samples are also more expensive to survey because cluster sampling is more difficult to achieve. Grade-based cohorts have been the dominant model in international studies of student achievement, but that pattern may not continue.

c) Construct comparisons

Most large-scale assessment studies focus on comparing results across countries or across Canada's various educational jurisdictions, therefore we felt rather comfortable comparing our studies' results from the various subgroups. However, comparisons among linguistic subpopulations can be limited by the assessment development and validation process, translation methods, and socio-cultural factors. The results of our studies led to the observation that teachers from French-language schools in Ontario and Quebec were similar in their teaching practices in reading and writing compared to those from English-language schools in Ontario. Given that education is a provincial jurisdiction in Canada, and thus teachers from English- and French-language schools in Ontario follow the same curriculum, use a common report card, and have their students all take the same provincial tests, we were very surprised by this observation. The reasons for these patterns are not obvious to us. They may be due to cultural differences in beliefs and practices, to different interpretations of questionnaire items, or to overall reactions toward the questionnaires (e.g., desirability).

Unfortunately, given the existing data sets, there was little we could do to investigate this issue further.

In the context of international assessments, theoretical construct or statistical comparability means that the factor structures being measured by various language versions of a test administered in different countries are invariant (Ercikan & Koh, 2005). However, in practice, differences in response patterns are seen across linguistic populations and countries (Black & Wiliam, 2005; Raveaud, 2004; Rémond, 2006). Such cultural differences can influence intrinsic interest and familiarity of the content of items (Rémond, 2006) and teaching practices.

Technical and Statistical Challenges

Technical challenges include a discussion about the nature of the data collected by large-scale assessments, scaling issues, missing data, and statistical data analysis models.

a) Nature of data

Given our interest in the teacher background questionnaires, and particularly in actual teaching and assessment practices, we noticed that the items generally requested information on level of importance attached to the statement offered, on frequency of use, or on quantity. There were very few questions on the quality of practice. In one instance, PIRLS inquired about the frequency of types of assessment used and included "other" as an option, but did not instruct teachers to identify these. Questions on the extent to which teachers agreed with various classroom practices—such as having students create collective novels, teaching explicit reading strategies, or encouraging meta-cognitive assessment strategies—would likely be more useful and tied to current theories of learning and assessments.

Similar limitations were previously echoed in another study of TIMSS-R's data. In Rodriguez's (1999) thesis on linking classroom assessment practices to large-scale test performance in mathematics using TIMSS's data, the author warned against limiting background questionnaire items to frequency. Rodriguez (1999) stressed the importance of developing items that would be indicators of quality to teaching and assessment practices. Another small-scale quantitative and qualitative study on the psychometric qualities of TIMSS's teacher background questionnaires from 1995, 1997, and 2001 demonstrated

that items on teaching practices focused principally on frequencies, were not sensitive to major variations in curriculum reform, and were subject to misinterpretations by teachers (Plouffe, 2007).

b) Scaling

The TIMSS-R 1999, PIRLS 2001, and SAIP 2002 teacher background questionnaires mainly utilized Likert-type items with scales offering two to four options. In our studies we found that ordinal scales posed specific qualitative and quantitative challenges. For example, the PIRLS 2001 used the scale "every day or almost every day," "once or twice a week," "once or twice a month," or "never or almost never" to determine how often the teachers had students do group projects related to what they read (Question 16, Item g). Although theoretically desirable, it is highly unlikely that in practice teachers do group projects every day, thus invalidating one or two of the four options. For other items, almost 100% of teachers opted for the most positive response, thus limiting variance in their answers and resulting in skewed distributions. These characteristics restrict access to appropriate statistical procedures and models.

Scaling is an issue that designers of large-scale assessment questionnaires should be conscious of because surveys generally result in nominal and ordinal level data. The background questionnaires structure their answers as either yes/no types of responses, or with limited options that may limit data analysis to exploratory exercises, non-parametric statistical analyses, and descriptions. Associations between variables can be assessed with, for example, chi-square analysis, but further examination of the relationships by using parametric statistics (such as structural equation modelling) among the variables is often not possible (Shepard et al., 1999). The use of constrained categorical scales, such as a 5-point ordinal scale, results in low variability in achievement scores and is not ideal for secondary analysis of correlationally based analysis, such as hierarchical linear models. Consequentially, it is recommended that continuous scales be adopted because they would most likely lead to more meaningful models of school performance through secondary analysis (Rogers et al., 2006).

c) Missing Data

Our investigation of the TIMSS-R 1999 study data required linking variables from three sets: student background questionnaire, teacher

background questionnaire and student results. Our second study looked at data from six subpopulations of teachers (three for the PIRLS data and three for SAIP). Throughout the secondary analyses, many cases were eliminated for various reasons. First, all cases with missing identification were removed. Second, in order to maintain anonymity within the minority language subpopulation, classes with fewer than 10 students were also eliminated from the analyses. Linking files also resulted in eliminating cases with large numbers of missing data. Table 1 shows the initial and resulting samples.

Table 1
Minority group sample sizes for the three large-scale assessments

Assessment	Initial sample	Resulting sample
TIMSS-R (1999)	73	36
PIRLS (2001)	138	80
SAIP (2002- Writing III)	182	76

Up to 42% of the data were missing from the teacher samples once cases were removed as a result of merging with student files. For the SAIP data, the total 76 teachers were actually representing two sub-groups: 31 of these taught 13-year-old students and 45 taught 16-year-old students. Given the initial relatively small sizes of minority groups, we hesitated to use extrapolation or modelling strategies to deal with missing data and instead conducted analyses with only full cases. We believed that doing otherwise would have distorted the data and would not have necessarily provided a better alternative, despite the large bias that may have resulted in treating files with large proportions of missing data.

Missing data is a significant problem for secondary analysis because it can reduce sample size to the point that most desirable statistical procedures are not possible (Rogers et al., 2006). Beaton and Robitaille (2002) suggest that having a quality assurance monitor for data collected would be helpful in achieving lower rates of incomplete or missing data. Rogers et al. (2006) agree that closer monitoring of instrument administration would ensure that the data are complete and reliable. However, monitoring will not resolve all missing data issues or even other assessment difficulties.

It appears that no matter how well a survey is designed and administered, response rates will be less than 100% and will vary from country to country. Three scenarios may explain these: nonresponse leading to data missing completely at random (MCAR), missing at

random (MAR), or non-ignorable missingness (Raudenbush & Kim, 2002). Solutions have been proposed for MCAR and MAR responses. For instance, data that are MCAR or MAR can justifiably be ignored during calibration and scoring (Dunn, Falenchuk, & Childs, 2005). However, numerous studies have shown that missing data, unless they are design-related (e.g., due to alternate test forms, targeted testing, or adaptive testing), usually cannot be considered as MCAR or MAR and generally lead to biased ability estimates (Dunn et al., 2005).

Although it is difficult to know why individual examinees fail to respond to items, failure to respond is in fact a behaviour, one that may reflect ability. Missing responses may contain some additional information that can be used to predict ability. Instead of guessing why an examinee failed to respond to an item and potentially biasing the ability estimates, nonresponse can be treated as a response category. Modelling the missing responses as a score category may offer an alternative for treating missing data. However, adequate performance models, such as hierarchical linear procedures, require large sample sizes at higher levels because they assume random sampling for each level.

d) Data analysis models

Studies that sample classes instead of students lend themselves to multilevel statistical models such as the hierarchical linear model (Bryk & Raudenbush, 1992; Wang 2001). This model allows estimation of variance due to sampling each level unit, student, class, school, and population. The relatively small samples in our studies could not be subjected to hierarchical linear models nor to structural equation modelling (SEM) mainly because of inadequate variables to sampling unit ratio and to lack of variance in most of the background data for each item. The HLM rule of thumb of 10 observations per predictor (Hoffman, 1997) did not apply to our data set. We were limited to the use of simple descriptive statistics, such as the use of exploratory or confirmatory factor analyses or contingency tables.

The object of school effectiveness research is to explore differences within and between schools by investigating the relationship between explanatory and outcome factors (Fox, 2004). This involves choosing an outcome variable, such as examination achievement, and studying differences among schools after adjusting for relevant background variables. A general acceptable statistical model in the assessment

of school effectiveness requires the deployment of multilevel analysis techniques (Fox, 2004). A multilevel sampling design reflects the structure of students nested in classes, and classes within schools and variance components are modelled at each sampling level where schools and classes are regarded as random effects (Fox, 2004). Furthermore, the model takes into account the homogeneity of results of individual pupils in the same school, since pupils in the same school share common experiences.

Moreover, large databases are typically multidimensional, and it is sometimes tempting to put all of the possible relevant variables into a multivariate statistical program to simply see what results are obtained (McCall & Appelbaum, 1991). Even when the initial set of variables have been carefully selected, too many variables can produce an over-determined set of predictors or outcomes, which may reduce the sensitivity of the analysis and make it more difficult to achieve significance and interpretability. When a large number of variables are used in a multivariate analysis, one stands to suffer from the infinite dilution problem (i.e., even a very strong effect seen on one or two variables can be completely lost by the inclusion of a substantial number of variables that do not reflect the effect [McCall & Appelbaum, 1991]).

Unless aggregated, databases from individual minority populations are usually relatively small and therefore do not lend themselves to multivariate or multilevel statistical models. However, the study of such populations is highly valuable and encouraged in order to ensure greater societal fairness and equity. This signals a need for greater attention to the development of secondary analysis approaches that take into account the specificities of such databases. The use of correspondence analysis or dual analysis may be some examples of statistical models to consider.

CONCLUSION

The personal account describing the challenges our research team faced in applying secondary analyses to minority group data sets from the TIMSS-R 1999 math, PIRLS 2001 reading, and SAIP 2002 writing III assessments is an important contribution to the field. In addition to reporting personal challenges, references to other studies provide a comprehensive view of the practical, conceptual, and technical/statistical hurdles that secondary analysis of minority group data present.

It is important to recognize that the nature of secondary data analysis is viewed as explanatory nonexperimental research (Rogers et al., 2006). The results indicate that the value of assessments may lie more in their potential for generating hypotheses about causal explanation than in their use as platforms for testing hypotheses. What can be expected from comparative studies are provocative new hypotheses about what may account for the differences in student achievement and not a cause-and-effect relationship (Porter & Gamoran, 2002).

Additionally, the practice of secondary analysis will vastly benefit from the integration of qualitative and quantitative data. The majority of current assessment data are quantitative; however, qualitative data can aid drastically in understanding the effects of context and culture on student achievement (Porter & Gamoran, 2002). New designs and new analysis strategies need to be created if the desired integration of quantitative and qualitative data is to be achieved. Qualitative data from small-scale studies hold promise for informing the direction of large-scale assessment work.

The literature emphasizes the need for collaborative multidisciplinary teams that comprise policy researchers, testing specialists, psychometricians, sampling specialists, data analysts, teachers, and principals (Rogers et al., 2006). Such individuals will enhance secondary data analysis by providing a network and access to knowledgeable persons to address concerns. They will also develop comprehensive designs allowing and focusing the secondary analyses. Moreover, they will ensure variables included are relevant to student learning, use instruments and scoring that will result in higher degree of variation, and sample representative and large enough populations to permit secondary data analysis (Porter & Gamoran, 2002). The focus on such collaboration should be to inform educational practice and improve student learning.

Rew and colleagues (2000) encouraged the use of secondary analysis of existing data sets as a reasonable alternative to conventional studies in spite of its limitations. Despite the problems we encountered and the process we experienced, we agree with them and also encourage the use of secondary analysis. Our reflections on the challenges we faced suggest to us, and we hope to other researchers, that attention to the concerns we have highlighted can minimize difficulties in using secondary analysis. By recognizing the problems that are occurring in the field, we can enhance our understanding and work toward possible improvements of secondary data analysis and assessments in

the future. As a whole, by better understanding the challenges facing educational statistics, we will be better equipped to explore complex issues and improve educational practice and policy.

ACKNOWLEDGEMENTS

We thank the Social Sciences and Humanities Research Council, Canada for three grants related to the studies reported in this article.

REFERENCES

Beaton, A.E., & Robitaille, D.F. (2002). *A look back at TIMSS: What have we learned about international studies?* In D.R. Robitaille & A.E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 408–418). Dordrecht, NL: Kluwer.

Black, P., & Wiliam, D. (2005). Lessons from around the world: How policies, politics and culture constrain and afford assessment practices. *The Curriculum Journal, 16*(2), 249–261.

Bouchamma, Y., & Lapointe, C. (2007, June). *Difficultés liées à l'étude des déterminants de la réussite scolaire des élèves du Canada francophones au PIRS écriture III 2002*. Conférence prononcée à l'atelier Les enquêtes comparatives du rendement scolaire et la question des langues officielles, Faculté d'éducation, Université d'Ottawa.

Brooks-Gunn, J., Phelps, E., & Elder, G.H. (1991). Studying lives through time: Secondary data analyses in developmental psychology. *Developmental Psychology, 27*(6), 899–910.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Clarke, S.P., & Cossette, S. (2000). Secondary analysis: Theoretical, methodological, and practical considerations. *Canadian Journal of Nursing Research, 32*(3), 109–129.

Crocker, R.K. (2002). *Learning outcomes: A critical review of the state of the field in Canada* [Submitted to the Canadian Education Statistics Council]. Retrieved December 2006 from <http://www.cmec.ca/stats/pcera/LearningOutcomes_StateoftheField_RCrocker2002.pdf>.

Crow, G.M., Levine, L., & Nager, N. (1992). Are three heads better than one? Reflections on doing collaborative interdisciplinary research. *American Educational Research Journal, 29*(4), 737–753.

Dunn, J., Falenchuk, O., & Childs, R.A. (2005, April). *Missing data and the nominal model: A possible solution?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23–25.

Fox, J.O. (2004). Applications of Multilevel IRT Modeling. *School Effectiveness and School Improvement, 15*(3–4), 261–280.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education, 11*(3), 319–330.

Gorard, S. (2002). The role of secondary data in combining methodological approaches. *Educational Review, 54*(3), 231–237.

Herry, Y. (2000). Enseignement et apprentissage des sciences : résultats de la troisième enquête internationale. *Revue des sciences de l'éducation, 26*(2), 347–366.

Hofferth, S.L. (2005). Secondary data analysis in family research. *Journal of Marriage and Family, 67,* 891–907.

Hoffman, D.A. (1997). An overview of the logic and rationale of hierarchical linear models—A special issue: Focus on Hierarchical Linear Modeling. *Journal of Management, 6*(23). <http://findarticles.com/p/articles"mi_m4256"is_n6_v23/al_20446317>.

Kiecolt, K.J., & Nathan, L.E. (1985). *Secondary analysis of survey data*. Beverly Hills, CA: Sage.

Kimpson, S.A. (2005). Stepping off the road: A researcher's story of challenging method/changing methodology. *McGill Journal of Education, 40*(1), 157–179.

Klinger, D. (2007, June). *Assessment cultures in Canadian education: The provincial perspectives*. Paper presented at the Large-scale Assess-

ments of Achievement and the Official Languages workshop, Faculty of Education, University of Ottawa.

Landry, R., & Allard, R. (2002). *Résultats pancanadiens des élèves francophones en milieu minoritaire aux évaluations du PIRS : variables contextuelles et conséquences pédagogiques*. Rapport de recherche soumis au Conseil des ministres de l'Éducation (Canada). Toronto: Conseil des ministres de l'Éducation.

Lévesque, É., & Bouchamma, Y. (2006, June). *Quelle supervision de l'enseignement pour améliorer la réussite des garçons et des filles : ce que nous suggèrent les résultats du Programme d'indicateurs de rendement scolaire (PIRS) écriture III 2002 PIRS*. Conférence prononcée à l'atelier Vers une conception partagée des évaluations à grande échelle en lecture et en écriture, Faculté d'éducation, Université d'Ottawa.

Levin, B. (2003). *Helping research in education to matter more*. Retrieved October 2003 from <http://www.sshrc.ca/web/whatsnew/initiatives/transformation/ben_levin.pdf>.

Maier, M.H. (1999). *The data game: Controversies in social science statistics* (3rd ed.). Arnmonk, NY: M.E. Sharpe.

Malone, K. (2006). Environmental education researchers as environmental activists. *Environmental Education Research, 12*, 375–389.

McCall, R.B., & Appelbaum, M.I. (1991) Some issues of conducting secondary analyses. *Developmental Psychology, 27*(6), 911–917.

Plouffe, S. (2007). *Analyse de la qualité psychométrique des items issus des questionnaires contextuels de la TEIMS*. Unpublished master's thesis, University of Ottawa.

Plouffe, S., Simon, M., & Loye, N. (2005, May). *Facteurs déterminants sur le rendement en mathématiques des élèves francophones minoritaires de l'Ontario aux études à grande échelle—Premiers résultats*. Conférence prononcée à la Société Canadienne pour l'Étude de l'Éducation (SCÉÉ), London, Ontario.

Porter, A.C., & Gamoran, A. (2002). *Progress and challenges for large-scale studies*. In National Research Council (Ed.), *Methodological advances in cross-national surveys of educational achievement* (pp. 3–23). Washington, DC: National Academy Press.

Raudenbush, S.W., & Kim, J. (2002). Statistical issues in analysis of international comparisons of educational achievement. In National Research Council (Ed.), *Methodological advances in cross-national surveys of educational achievement* (pp. 267–294). Washington, DC: National Academy Press.

Raveaud, M. (2004). Assessment in French and English infant schools: Assessing the work, the child or the culture? *Assessment in Education, 11*(2), 193–210.

Rémond, M. (2006). Éclairage des évaluations internationales PIRLS et PISA sur les élèves français. *Revue française de pédagogie, 157,* 71–84.

Rew, L., Koniak-Griffin, D., Lewis, M., Miles, M., & O'Sullivan, A. (2000). Secondary data anlysis: New perspective for adolescent research. *Nursing Outlook, 48,* 223–229.

Robitaille, D.R., & Beaton, A.E. (2002). *Secondary analysis of the TIMSS data*. Dordrecht, NL: Kluwer.

Rodriguez, M.C. (1999). *Linking classroom assessment practices to large-scale test performances*. *Dissertation Abstracts International, 60*(10A), 3638.

Rogers, T.W., Anderson, J., Klinger, D.A., & Dawber, T. (2006). Pitfalls and potential of secondary data analysis of the Council of Ministers of Education, Canada National Assessments. *Canadian Journal of Education, 29*(3), 757–770.

Savard, D., Sévigny, S., & Beaudoin, I. (2007). Évaluations à grande échelle de l'écriture : lien positif entre le score holistique et les composantes de l'écriture. *Canadian Journal of Program Evaluation, 22*(3), 99–119.

Shepard, M.P., Carroll, R.M., Mahon, M.M., Moriarty, H.L., Feetham, S.L., Deatrick, J.A., et al. (1999). Conceptual and pragmatic considerations in conducting a secondary analysis. *Western Journal of Nursing Research, 21*(2), 154–167.

Simon, M., Turcotte, C., Ferne, T., & Forgette-Giroux, R. (2007). Pratiques pédagogiques dans les écoles de langue française de l'Ontario selon les données contextuelles du PIRLS 2001. *Mesure et évaluation en éducation, 30*(3), 59–80.

Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher, 30*(6), 17–21.

Marielle Simon has been a member of the Faculty of Education of the University of Ottawa since 1994. She is a full professor, and her research interests are in measurement, evaluation, and assessment. She has obtained various SSHRC grants in the area of secondary analyses of large-scale assessments data and of classroom assessments.

Nicole Roberts has just completed her Masters in Educational Counselling at the University of Ottawa. She worked as a research assistant on some of the SSHRC-funded studies directed by Marielle Simon and Renée Forgette-Giroux.

Robin Tierney is a Ph.D. student in the Faculty of Education of the University of Ottawa. She has taught language arts in elementary schools in the Ottawa area. She studies in student assessment.

Renée Forgette-Giroux has recently retired from the Faculty of Education of the University of Ottawa. She was a full professor in Measurement and Evaluation. She is still very active on some of the SSHRC-funded studies and student supervision.