A Framework for Using Consequential Validity Evidence in Evaluating Large-Scale Writing Assessments: A Canadian Study

David H. Slomp
University of Lethbridge

Julie A. Corrigan
University of Ottawa

Tamiko Sugimoto
University of Lethbridge

The increasing diversity of students in contemporary classrooms and the concomitant increase in large-scale testing programs highlight the importance of developing writing assessment programs that are sensitive to the challenges of assessing diverse populations. To this end, this paper provides a framework for conducting consequential validity research on large-scale writing assessment programs. It illustrates this validity model through a series of instrumental case studies drawing on the research literature conducted on writing assessment programs in Canada. We derived the cases from a systematic review of the literature published between January 2000 and December 2012 that directly examined the consequences of large-scale writing assessment on writing instruction in Canadian schools. We also conducted a systematic review of the publicly available documentation published on Canadian provincial and territorial government websites that discussed the purposes and uses of their large-scale writing assessment programs. We argue that this model of constructing consequential validity research provides researchers, test developers, and test users with a clearer, more systematic approach to examining the effects of assessment on diverse populations of students. We also argue that this model will enable the development of stronger, more integrated validity arguments.

A defining characteristic of Canadian identity is diversity, and the multicultural and multiethnic constitution of the Canadian population is as diverse as the country's geographic regions. Aboriginal peoples, for example, are the fastest-growing population in Canada with their youth population growing more than 20% between 2006 and 2011 (Statistics Canada, 2011a). Canadians report using more than 200 different languages as mother tongues, and one in five Canadians speaks a language other than French or English at home (Statistics Canada, 2011b). Canada also has the highest percentage of foreign-born residents among the Group of Eight (G8), and almost one in five people living in Canada is a visible minority (Statistics Canada, 2011c). Canadians are duly proud of being the first country in the world to adopt multiculturalism as an official policy (Citizenship and Immigration Canada, n.d.), often boasting of the merits of the cultural mosaic over

the melting pot approach taken by other countries. At a time when, as McLuhan (1962) might say, the classroom has become a global village—suffused with cultural, linguistic, socioeconomic, and geographic differences—demographic diversity is rapidly changing the nature of Canadian schools.

Increases in diverse student populations in the Canadian educational system have been accompanied by a concomitant proliferation of large-scale testing. On the surface, these two movements (one toward increased diversity, the other toward increased standardization) seem at odds with one another. The resulting tension highlights the need for a systematic approach to understanding the consequences that accrue as a result of tests, especially with respect to their effects on diverse populations of students. Bearing this in mind, researchers require measurement procedures sensitive to the differential impact of assessment practices on diverse populations. The question of how to measure these effects remains a problem, however, as the field lacks a systematic approach to collecting consequential validity evidence (Lane, 2013). Building on the work of Messick (1989) and Kane (2006)—along with the emerging work of White, Elliot, and Peckham (in press)—we offer an articulated model of validation for educational assessments.

The goal of this paper is to present a conceptual framework for gathering consequential validity evidence and to use this framework within the context of large-scale, government-mandated writing assessment programs in Canada. In the following article, we articulate a model—grounded in the conceptual systems articulated by Cronbach and Meehl (1995), Messick (1989), and Kane (2013)—for collecting consequential validity evidence in relation to writing assessment; then, we use the model to examine the consequential validity research on Canadian large-scale writing assessment; and finally, we discuss the lessons we learned after examining consequential validity from this perspective.

Context: Historicizing Validity

Historically, validity has been characterized as providing "information [indicating] the degree to which a test is capable of accomplishing certain aims" (American Educational Research Association & National Council on Measurements Used in Education, 1955, p. 15). In the past, validity arguments relied primarily on content, concurrent, or predictive validity evidence. Construct validity evidence was called upon only when these three forms of validation had failed. Messick's (1989) formulation, however, has shifted construct representation from the periphery to the center of validity theory. In that formulation, Kane explains, Messick argued for a unified theory of validity dependent on construct integrity and one that gives "the consequential basis for validity equal billing with the evidential basis" (2006, p. 21).

Yet, the construct model has opened up a seemingly endless process of validation. In response to this problem, Kane (2006) has developed an argument-based approach to validation, one that maintains construct validity as a central element but that more clearly defines the kinds and extents of evidence required to support test validation. At the same time, Kane's model recognizes that the process of validation is never complete; evolution of theories, shifts in educational and

testing contexts, and critiques of current practices all necessitate a recursive process of validation.

Kane's model (2006, 2013) describes a two-stage, argument-based approach to validation, one that begins with an articulation of an interpretation/use argument that outlines all of the claims that are predicated on a test's scores, such as "the network of inferences and assumptions inherent in the proposed interpretation and use" (Kane, 2013, p. 2), followed by a program of research that tests the warrants for those claims. While this approach seems reasonable from a construct validity perspective, it is more problematic from a consequential validity perspective because the consequences of test use and interpretation often extend beyond the scope of the intended interpretation and use of the test.

Consequential Validity Evidence

The role of consequential validity evidence within a broader validity framework has been much debated over the years (Borsboom, Mellenbergh, & van Heerden, 2004; Cizek, Bowen, & Church, 2010; Maguire, Hattie, & Haig, 1994; Popham, 1997, 1999). This debate has been protracted over two main questions: First, should consequences be considered within the validity framework itself? And second, to what extent can test developers and users be held accountable for the unintended consequences accruing from test use?

Messick's (1989) work has become a lightning rod within this debate because he fuses construct validity and consequential validity issues into one unified theory. Kane (2013) supports Messick's position that unintended consequences should be included within the validity framework, and that consideration of those unintended consequences should be limited to three areas: (1) those with a potential for substantial impact in the population (or subpopulations) of interest; (2) those with particularly adverse impacts; and (3) those with systemic consequences.

While the most frequent critique of Messick's work focuses on limiting the scope of consequences for which test users and developers must be held responsible (Borsboom, Mellenbergh, & van Heerden, 2004; Popham, 1997, 1999), Cronbach (1988) and Inoue (2009) argue the opposite: Messick's formula unduly limits the scope of consequences that can be considered pertinent to those issues that result only from flaws in construct representation. Cronbach argues that if the negative consequences accruing from test use are severe enough, they can on their own provide sufficient cause to discontinue the use of the test regardless of their connection to construct validity concerns. Inoue favors Cronbach's broader perspective while critiquing it for its lack of a systematic focus on the role that power dynamics and sociopolitical histories play in shaping a test's impact on populations of test-takers. To address this problem, he proposes adopting "racial validity" as an additional line of investigation within the current validity framework. While Inoue's rationale has merit, a more productive approach to addressing this problem is to better define a systematic approach to collecting and weighing consequential validity evidence, one that is sensitive to racial and sociocultural realities (Solano-Flores, 2011) and that examines both intended and unintended outcomes.

A Model for Collecting and Evaluating Consequential Validity Evidence

What follows in this article is an explication of a systematic approach to collecting and integrating consequential validity evidence. We follow this with an illustration of this model constructed through a series of instrumental case studies (Stake, 2005). We have chosen to develop a model derived from a study of Canadian writing assessment because the diverse population of students in Canadian schools is reflective of the diversity in many other developed nations and because it is this issue of student diversity that highlights the need for consequential validity research.

Our model for collecting consequential validity evidence appears in Figure 1. This figure is derived both from Kane's (2006, 2013) model for constructing a validity argument and from White, Elliot, and Peckham's (in press) extrapolation of Kane's model. The main modification we have made to these models is to infuse consequential validity considerations into each stage of the design and validation process. This modification, we believe, brings Kane's model into greater alignment with Messick's construct- and consequence-focused conception of validity.

At the heart of this model is the recognition that decisions at every step of the assessment process carry both intended and unintended consequences. Following from this foundational assertion, our model suggests that those who design and use tests have an obligation to examine both the intended and unintended consequences that accrue as a result of their decision-making process and, where warranted, to remedy negative unintended consequences. The burden this process places on assessment developers and users is significant. And so, from the outset, one key limitation of this claim needs to be stated: the expectations for collecting consequential validity evidence should be proportional both to the stakes attached to the assessment program and to the complexity of the construct under investigation (Koretz & Hamilton, 2006; Nichols & Williams, 2009).

In this model, we define the following sources of evidence connected to key consequential validity questions (corresponding with the octagons in Figure 1): construct definition, construct irrelevant variance, design process, scoring procedures, sampling plan design, disaggregated performance, construct remodeling, and implications (intended and unintended). At each of these stages, our reconceptualization adds a series of construct and consequential validity questions (see Table 1) that provide focus to a program of consequential validity research.

Construct Definition

The more complex the construct being assessed, the less likely it is that we can ever capture that construct completely or exhaustively. This limitation particularly applies to constructs such as writing ability. On the one hand, our understanding of the construct itself is continually changing in response to ongoing research and shifts in theoretical perspectives (Camp, 2012; Yancey, 2009). On the other hand, the contexts, modalities, and technologies for writing are continually expanding, changing the nature of the writing construct in the process. The key issue with respect to this category of test design and validation is to determine both how well

the construct being measured is understood and how solid the consensus regarding that construct definition is. Even consensus within the field should not be taken as a reason for undue confidence, however. Complex constructs such as IQ, for example, once seemed clearly established, only to be significantly problematized later (e.g., Gould, 1996; Schönemann, 1997). The more tentative test developers are at this early stage, the more likely they are to make careful decisions about test design and use.

Construct Irrelevant Variance

Further adding to the need for tentativeness in the design of tests and the use of test scores is the challenge posed by construct irrelevant variance. In his validity model, Kane (2006) identifies three potential sources of construct irrelevant variance that can infiltrate the target domain being measured: observation methods, context,

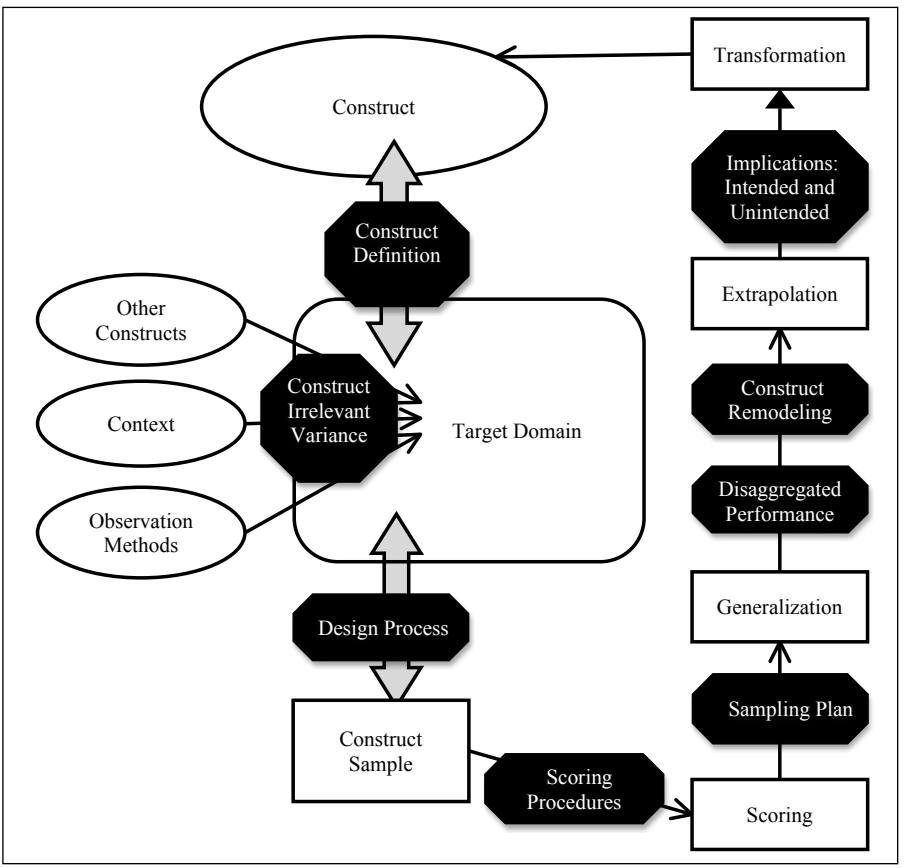

FIGURE 1. ***Extrapolation of figures from both Kane (2006) and White, Elliot, and Peckham (in press). Black octagonal boxes represent sources of consequential validity evidence.***

Table 1. **Consequential Validity Questions by Source**

Sources	Validity Questions	Consequential Validity Questions
Construct Definition	Has the construct been specifically defined?	How well is the construct understood? How stable is this construct across social, cultural or racial contexts?
Construct Irrelevant Variance	To what extent do other constructs, contexts, and methods of observation introduce the potential for construct irrelevant variance into the measurement process?	Are related yet distinct constructs interfering with the construct definition as it is embedded in the construct sample?
Design Process	Is the assessment designed to ensure that the construct is measured effectively and that all populations impacted by the assessment are considered before the assessment is launched?	Does the assessment design contribute to potentially adverse impacts, impact on populations demonstrated to be at-risk, and educational systems serving those students?
Scoring Procedures	Do scoring procedures adhere to best practices for achieving consistency while also supporting construct validity?	How do scoring procedures influence assessment outcomes, student populations and the educational systems serving those students?
Sampling Plan	Does the sampling plan identify diverse populations who might be differentially impacted by the assessment?	Does the sampling plan ensure that each population is represented in sufficient quantity to allow descriptive and inferential analysis? If not, what justification is provided for limiting the sampling plan?
Disaggregated Performance	Does the design allow for each population to be examined for writing performance?	Can differences in performance between all populations be attributed to actual differences in ability in relation to the construct being measured?
Construct Remodeling	Does an examination of student response processes demonstrate that the writing task is measuring the same construct for different populations of students?	Do differences in student response processes lead both to inaccurate ratings of their performance and to improper decisions based on those ratings?
Implications: Intended	Taken collectively, does the evidence gathered indicate that the assessment has achieved the purposes or goals for which it was designed?	What are the intended consequences both for each population impacted by the assessment, and for the educational systems serving those students?
Implications: Unintended	Taken collectively, does the evidence provide an understanding of unintended impact, whether positive, negative, or unknown?	What are the unanticipated positive, negative, and unknown consequences both for each population impacted by the assessment, and for the educational systems serving those students?

and other constructs. First, it has long been understood by researchers that no method of observation is neutral in terms of its effect on the phenomenon under investigation. Second, the measurement context, as well, can introduce potential sources of irrelevant variance. Understanding the influence that testing context exerts on test takers is important because this helps us to better understand what test scores represent for particular individuals or populations. The third source of irrelevant variance generally subsumes the previous two sources in that the problem associated with choices of measurement tools or with defining the measurement context is that these introduce other associated constructs into the target domain.

Design Process

The design process begins with the identification of aims and purposes for the assessment tool being developed—for example: diagnostic information to assist in individual student program planning; formative assessment data to help teachers plan instruction; summative assessment data to inform students regarding levels of achievement they have attained; and achievement data for teacher-, program-, institution-, or system-level accountability purposes. Once the purpose of the assessment tool has been determined, stakeholders need to be brought into the design process. Historically, this has not been the case, and when other stakeholders are consulted, their contributions tend to be subsumed within the dominant paradigm established by the assessment specialists (Huot, 2002). From a consequential validity perspective, a more inclusive design process[1] should lead to increased support from a range of stakeholder groups, reduce the potential for bias, and generate more appropriate assessment decisions. Different stakeholders, too, are likely to be more in tune with the range of potential unintended consequences that accrue as a result of design decisions. Broad (2000) suggests that publicizing the design and evaluative decision-making process (including points of both agreement and dissension among stakeholders) is important to enhancing transparency and ultimately the validity of writing assessment tools.

Scoring Procedures

Historically, consequential validity concerns related to scoring procedures have focused on the issue of interrater reliability. The validity concern here is that a student's score should not be a function of who scored the test, but rather a reflection of a student's performance in relation to the construct measured, as expressed through the scoring criteria. While the need for consistency in scoring is well accepted, the field of writing assessment has engaged in a lengthy debate about both the costs to validity of achieving high degrees of reliability and how best to measure and achieve reliability (Lynne, 2004; Moss, 1994; Parkes, 2013). The consensus within the field is that these two values exist within a state of dynamic tension (Slomp & Fuite, 2005). Kane (2013) observes that reliability is often enhanced through greater standardization. He cautions, though, that decisions that attempt to increase reliability need to be balanced against the validity costs associated with them, stating that "standardization of any aspect of the testing procedure that is not also fixed in the target domain introduces a source of systematic error" (p. 30) linked to the

problem of construct underrepresentation. One of the primary methods of achieving greater interrater reliability has been to purposefully design rubrics that allow for high degrees of consistency in scoring. In such cases, matters of organization, correctness, and choice are often emphasized because these criteria can be scored more consistently. More complex aspects of the construct, such as creativity, critical thinking, or metacognition, however, are underrepresented by many rubrics. Hillocks (2002) illustrates this problem in his study of state writing assessment programs, drawing a clear link between construct-criterion alignment and broader social consequences when he directly attributes students' development of limited thinking skills to a testing program that measures truncated thinking rather than the full development of ideas. His observations highlight the importance of not only considering measures of consistency when evaluating scoring procedures, but also examining the procedures themselves with respect to construct alignment.

Sampling Plan

Once scores have been generated, decisions based on those scores need to be made. Of particular importance to the issue of assessing diverse populations of students is the need to collect clear data on the range of populations who will be affected by the test. Many testing agencies develop sampling plans that focus on reporting school, school district, or other forms of geographic data. Greater attention needs to be paid, however, to identifying and effectively sampling populations of test takers who may be differentially affected by test scores. Without an adequate sampling plan in place, it is difficult to determine the extent to which this difference in experience translates into differences in test scores for various populations of students.

Disaggregated Performance

Sampling plan decisions, then, impinge directly on assessment users' capacity to disaggregate student performance in meaningful ways. Disaggregation by population is important from a consequential validity perspective because it enables test users to better understand whether any prior design decisions have had a differential impact on subgroups of students. If so, this information can point to the need for assessment redesign. This information is also important for helping test users understand whether the decisions they make based on test scores are warranted with respect to the construct being measured. Does a placement test, for example, enable decisions based solely on student performance in relation to the construct being measured, or do these decisions reflect other factors related more to issues of culture, race, or socioeconomic status? Detailed sampling plans linked to disaggregation of data provide confidence both to generalization inferences and to the decisions about test-takers (both individually and collectively) that are based on those inferences.

Construct Remodeling

The process of disaggregation can reveal differences in performance of diverse populations, but this process does not always explain why these differences occur. It does, however, provide researchers with a starting point for answering this ques-

tion. Once differences in performances across populations have been identified, an examination of student response processes across this range of populations can be conducted. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) indicate that an examination of response processes can help test developers and users understand "which capabilities irrelevant or ancillary to the construct may be differentially influencing [each population's] performance" (p. 12). Methods for collecting this type of evidence include think-aloud protocols, focus group discussions, and electronic monitoring of student writing processes. Developing clearer understandings of the response processes of different groups helps to better explain variance in performance across groups, which in turn helps validity researchers to better understand whether test scores actually mean the same thing across different populations. From a consequential validity perspective, this is important because it enables test users to make better decisions about the meaning and use of test scores. If a writing test, for example, is shown to be measuring language proficiency instead of writing ability for certain populations, test users need to rethink whether those scores can be used for placement purposes in a writing program.

Implications, Intended and Unintended

Cizek, Bowen, and Church (2010) argue that one reason to reject consequential validity evidence as part of a broader validity argument is that this evidence is "'noncompensatory': that is, it cannot logically be combined with other forms of validity evidence 'into a coherent, integrated evaluation'" (p. 740). We could not disagree more. What has been lacking historically has been a systematic approach to integrating the range of data sources involved in developing a robust validity argument that includes consequential validity evidence to help answer questions such as the following: Taken collectively, does the evidence gathered so far indicate that the assessment has achieved the purpose or goals for which it was designed? If so, what effect does this have on the educational system (at the intended level) for which the test was designed? If not, how does this failure contribute to (a) particularly adverse impacts, (b) impacts on populations, and (c) impacts on relevant educational systems?

In the section that follows, we will illustrate the range of consequential validity evidence that can be collected within this framework. The primary method of ensuring an integration of validity evidence is through a two-stage process, one that leads, at each stage of this model, with a construct validity question, which is then paired with a consequential validity question (see Table 1). This process ensures that the focus of data collection and interpretation is constantly grounded in the integrity with which the test measures the construct it was designed to measure. What we have found particularly interesting about this range of data is how well both the quantitative and qualitative data across multiple studies agree with one another. We are not surprised by this, though, as mixed-methods researchers have been developing approaches to combining diverse data sets for some time now (Bryman, 2006; Denzin, 2010; Onwuegbuzie & Leech, 2005).

Case Studies within the Canadian Context

The impetus for us to design the model described in the first half of this article was to understand the extent to which agencies responsible for developing and administering large-scale writing tests in each province and territory in Canada were collecting and reporting consequential validity evidence. We also wanted to determine the current state of consequential validity research in Canada relative to these same testing programs.

In developing our approach we recognized that, because of the lack of consensus regarding the concept of consequential validity, a systematic method for collecting this type of evidence and incorporating it into a broader validity argument has not been agreed upon. Given this context, we expected that the consequential validity research conducted in Canada (or in any other jurisdiction, for that matter) would necessarily be both uneven and incomplete. For this reason, we report data using an instrumental case study approach. Stake (1995, 2005) defines instrumental case studies as a form of investigation in which the researcher's primary concern is what each case reveals about the phenomenon of interest, rather than the intricacies of each case itself. Within this methodology, the case "plays a supporting role, and it facilitates our understanding of something else" (Stake, 2005, p. 445).

Data Collection and Analysis

Our case studies were derived from two sets of documents: (a) government reports posted to Ministry of Education websites for each province and territory in Canada, and (b) peer-reviewed research reporting consequential validity evidence related to these same government-mandated assessments.

Government Reports

A systematic Internet search was performed for each of Canada's ten provincial and three territorial Ministry of Education government websites to collect any documents that would help answer the following questions: (1) What data or arguments are being used to justify/provide indicators of the impact of literacy assessments on teaching, learning, and achievement? (2) How are assessment results being used?

We chose to focus on publicly available documentation for this study because we agree with Brennan (2006) that testing agencies have an obligation to publicly present validity evidence in a timely manner. These documents are reliable self-reports prepared by teams of psychometricians and, as such, constitute excellent sources of information with respect to a test's impact in local contexts. In order to provide the most comprehensive results, a broad net was cast, allowing for a methodical search of all listed documents, links, brochures, graphs, and PDF files listed on the Ministry websites. The search did not exclude any documents, with the exception of those pertaining to adult and continuing education. Beginning on the Ministry home page, the search was conducted in a systematic progression of opening documents and links, and noting relevant content. In order to maintain structure, the home page would always be the returning point of reference. Once all links and documents were exhausted, the keyword search box was also utilized to ensure no documents were overlooked. In total, a comprehensive search for

documents posted on the Ministry of Education websites took approximately 120 hours. This search generated 64 documents that addressed our first question and 35 documents that addressed our second question.

Peer Reviewed Research

Next, our review of the literature focused on capturing the research on how large-scale writing tests influence education in Canada. We queried three databases, including ERIC (Education Resources Information Center), Academic Search Complete, and CBCA (Canadian Business and Current Affairs) Complete; the latter was added to ensure the inclusion of articles from a Canadian perspective. In addition to this, we searched the archives of four prominent Canadian education journals, including the *Canadian Journal of Education, McGill Journal of Education, Alberta Journal of Educational Research*, and *English Quarterly Canada.* In total, 33 searches were completed, resulting in 908 hits. We delimited the search to peer-reviewed journal articles from the years 2000 to 2012. Search terms included "standard* literacy test* Canad*" and "standard* assess*," in addition to "standard* test*," in combination with the names of all ten Canadian provinces and three territories. Of the 908 hits, only 19 articles were retained, based on the following criteria: (a) the article reported original research; (b) it was based on Canadian large-scale writing tests; (c) it reported on large-scale writing assessment results; and (d) it discussed consequential validity evidence.

These documents were then grouped by the type of validity evidence they provided (according to the model described in the previous section). Given the current state of consequential validity research in Canada, we did not anticipate that any one study or report would capture the full range of validity evidence sources presented in our model. Next, we chose the best exemplar document(s) from each group as case studies illustrating the types of consequential validity evidence that can and should be collected within our model. Our search methods enabled us to gather a comprehensive collection of source documents. This, in turn, enabled us both to select the strongest case studies for discussion and to speak with greater confidence about the state of consequential validity research in this context.

Illustration of Consequential Validity Evidence by Source

In the following subsections, we present instrumental case studies that illustrate the questions asked and types of data collected for each of the data sources we've articulated in our model.

Construct Definition

The testing agencies in Canada define the writing construct their tests are designed to measure with varying degrees of detail. Ontario's testing agency (Education Quality and Accountability Office, 2011) defines writing as measured on the Ontario Secondary School Literacy Test (OSSLT) as involving the following three traits:

- Writing skill 1: developing a main idea with sufficient supporting details
- Writing skill 2: organizing information and ideas in a coherent manner

- Writing skill 3: using conventions (i.e., spelling, grammar, punctuation) in a manner that does not distract from clear communication

The British Columbia Ministry of Education (n.d.) offers a more detailed, though very dated, construct definition (based on an article published in *SLATE* in 1979) for its grade 10 English examination, one that includes the following traits: the ability to develop ideas, structure and organize text, choose mode of discourse, develop appropriate tone and form, and appeal to possible audiences.

These two exemplars suggest that little was done to develop and critically interrogate the construct definitions that underpin these assessment tools. In contrast to the 13-trait model discussed by Elliot and Klobucar (2013), or the transfer-oriented model articulated by Beaufort (2007), these construct definitions seem simplistic, dated, and limited. These suspicions are supported by Peterson, McClay, and Main (2012), who examined the constructs for large-scale writing assessments administered at the grades 5–8 level in each provincial and territorial jurisdiction in Canada. In their study, Peterson, McClay, and Main defined writing ability according to two contemporary theoretical frameworks: a process-oriented theory of writing and a multiliteracies theory. They then conducted a deductive analysis of administrative documents, scoring guides, and exemplars for each provincial and territorial writing test to see how well each captured these construct features. They found that these exams did measure key features of the process-oriented aspects of the construct. They also found that these exams suffered from many issues of construct underrepresentation and construct irrelevant variance including the following: exams did not allow for a recursive writing process, nor did they measure students' ability to make choices related to topic, genre, purpose, or audience; only one exam context allowed students to choose the topics they were writing about; only two exam contexts provided students with an opportunity to respond to or receive feedback on their writing in process; only two exam contexts allowed students to work on their writing in class over the course of the year; only one exam context measured students' ability to compose across multiple modalities. In addition to these issues, the authors found only token acknowledgement of linguistic and cultural diversity, as well as Aboriginal culture, in writing prompts or support materials, with accommodations only for English language learners (ELLs).

Peterson et al. (2012) link these issues of construct underrepresentation to a range of potential consequential validity issues, specifically that "large-scale assessments are likely to become increasingly removed from the actual literacy practices of literate people" (p. 440) and that this growing divide between real-world literacies and school literacies "leads to the irrelevance of school literacy in the eyes of the young" (p. 440).

Construct Irrelevant Variance

The major sources of construct irrelevant variance are contexts, methods of observation, and other constructs. Regardless of the source, the primary consequence of irrelevant variance is that it distorts the picture of student ability that is being measured.

In respect to the issue of context introducing construct irrelevant variance into test scores, Klinger, Rogers, Anderson, Poth, and Calman (2006) studied the contextual and school factors associated with achievement on the OSSLT. In this study, they examined achievement score data, demographic information, and literacy activity data from 160,491 students using a process of hierarchical linear modeling to identify the relationship between student performance on the OSSLT and 12 student-level and 2 school-level variables. They found that

> the majority of the variability for both the reading and writing achievement scores was between students rather than between schools . . . [but] that less than 30 per cent of the student-level variability in achievement was accounted for by a set of available student variables. . . . Literacy related variables had small associations with both reading and writing. (p. 790)

They conclude their analysis stating that "such findings are discouraging from a policy perspective because they indicate that student contextual variables continue to have the most identified influence on achievement" (p. 790). They are careful to note that this degree of unexplained variability might be associated with limitations in the data set, including limitations in the number of variables available for examination. This high degree of unexplained variability, however, makes clear both the importance of context in shaping student performance and the danger of not adequately understanding the sources of that variance.

Methods of observation also can contribute to construct and consequential validity problems for an assessment program. Slomp (2008), for example, examined Alberta's grade 12 academic English exam—a timed-impromptu model of writing assessment—to determine how its design might introduce construct validity issues into the measurement context. In developing a construct model for this exam, Slomp analyzed the technical report for the test, the test itself, its scoring guides, and the information bulletin on the test. He found that the test was designed to measure the following traits: knowledge about language structure; knowledge about language as a communication tool; knowledge about the creation of voice; ability to generate, organize, and effectively present ideas within tightly controlled time frames; and ability to work effectively under pressure. Based on this analysis, he argues that because the exam measures students' ability to write under pressure, to generate ideas quickly, and to create polished first-draft writing, its design ensures that construct irrelevant variance is negatively impacting test scores. Slomp (2005) also found that these construct issues had a negative impact on student learning. Student participants in this study learned to value in their own writing the limited approach to process required for success on the exam rather than the more robust process described in the process literature.

Design Process

Openness in the design process is important to the integrity of an assessment program. In Canada, stakeholders in the education system are consulted through commissions, task forces, or public hearings. These discussions tend to focus on

broad educational policy and direction. With respect to the design of specific assessment programs, however, the level of stakeholder engagement tends to be much narrower. The one group often included in this process is teachers. In Ontario, educator perspectives are taken into account when operational items are selected "to ensure that they reflect the blueprint for the assessment and are balanced for aspects such as subject content, gender representations and provincial demographics (e.g., urban or rural, north or south)" (Education Quality and Accountability Office, 2011, p. 3). Additionally, the Education Quality and Accountability Office (EQAO) has formed a Sensitivity Committee composed of 24 educators who "provide expert advice from a specialized equity perspective to ensure that assessment materials are fair for a wide range of students" (p. 6). Their work informs the year-to-year development of EQAO tests.

In our review of the literature, we could find no studies that examined the consequences accruing from the design process.

Scoring Procedures

Provincial testing agencies in Canada pay close attention to ensuring that their scoring procedures generate high degrees of reliability. In Ontario, for example, scorers are trained to use the EQAO rubrics and anchor papers to generate consistent and appropriate grades for students' written responses. Once trained, scorers must pass a qualifying test before being permitted to work as members of the grading team. Once grading begins, each paper is scored independently by two graders. If their scores do not agree with one another, an expert grader is asked to make a final judgment on the paper. Additionally, graders are expected to assess 10 validity papers each day, maintaining a high degree of agreement with the expert graders across these 10 papers. Graders who fail to do so are retrained or dismissed. Through this process, the EQAO achieves 98% exact-adjacent agreement between scorers (scores between pairs of raters that were either identical or separated by one point on the rubric). Achieving this degree of consistency is an important accomplishment for this testing agency. But interrater reliability is only part of the picture when it comes to evaluating the effectiveness of scoring procedures. Test designers also need to consider the extent to which these scoring procedures compromise the test's construct validity.

In the research we reviewed, the EQAO has been criticized because its scoring criteria for the writing portions of the Ontario Secondary School Literacy Test do not adequately reflect the constructs intended. Ricci (2004) conducted a case study of one Ontario school, examining the influence that the OSSLT exerted on the school's curriculum, students, and teachers. His case study involved interviews with teachers, observations of classroom practices, and the collection and analysis of EQAO documents, school board policies, and school communications. Based on his analysis of these data, he reports that the school's response to the OSSLT was both to narrow the grade 9 and 10 literacy curriculum to those aspects of literacy being measured by the test, and to focus considerable classroom attention on helping students develop appropriate test-taking skills. This decision, he reports, was largely a consequence of the school's concern that the scoring procedures for

the OSSLT introduced construct irrelevant variance into the testing situation. He reports:

> Based on the results compiled by EQAO, the school offered advice to their future test takers informing them of the areas on which first-year test takers faired [*sic*] poorly. The primary focus was not literacy, but test taking. The instructions that teachers received were "Explain [to students] that the sessions planned for the next few weeks are NOT designed to make students literate! . . . Emphasize that these . . . sessions ARE designed to IMPROVE THEIR TEST-TAKING SKILL. We know from the detailed feedback we received from EQAO that even students who were apparently making a sincere effort last year lost points in a number of ways that had more to do with PROCEDURE than actual LITERACY." (p. 352; emphasis in the original)

The school's findings are supported by Fox and Cheng (2007), who similarly found that especially for L2 students, the OSSLT was measuring test-taking skills in addition to the literacy construct it was designed to measure. Ricci's (2004) findings highlight the consequential validity issues that arise when scoring procedures are understood to be undermining the construct validity of a writing exam.

Sampling Plan

Little detail is provided by any province or territory about the sampling plan used during the test development process. In the province of Nova Scotia, for example, it has been reported that a random sample of students is used for field testing (Nova Scotia Department of Education, 2010). A more complex procedure, stratified random sampling (otherwise known as proportional or quota sampling), involves dividing a population into homogenous subgroups and then taking a random sample from each. These subgroups might include minority groups such as ELLs, Aboriginals, students with learning disabilities, students from various socioeconomic classes, and so forth. It is important that these subgroups are represented during field testing, for if two subgroups that normally perform at the same level exhibit a statistically significant difference in performance on a particular test, it may be that the test is biased or that it is undermined by issues of construct-irrelevant variance.

For the actual assessment, all of the provinces and territories endeavored to include the entire population in their large-scale testing programs; however, in each instance, there was a small percentage of students who were absent or exempted. For example, in 2012, 93% of the grade 10 student population took the OSSLT, while 2% were absent and 5% deferred (Education Quality and Accountability Office, 2012). These percentages, presumably, do not include the students who had dropped out of school or were exempted.

The problems associated with limited sampling plans during the development phase are illustrated by Roos et al. (2006), who conducted a longitudinal analysis of population data (birth, school enrollment, diploma exam performance) for the cohort of Manitoba students born in 1984. Based on their analysis of these data, they argue that students who are not taking the test are just as important as

those who are. Those not taking the test—perhaps due to truancy, exemption, or having dropped out of school—are typically from a lower socioeconomic demographic. Since those who do not take the test do not have a test score, "test[s] will overestimate the performance of groups at risk for poor outcomes and provide distorted, inaccurate comparisons of school performance" (p. 698). Furthermore, as test performance often drives funding decisions and allocation of resources, districts and schools whose students are of lower socioeconomic status are typically disadvantaged when policy makers fail to consider those who were absent—not simply those who were present—on test day. More robust sampling plans that better consider diverse groups within the larger sample will help mitigate the limitations imposed by underrepresentation of populations.

Disaggregated Performance

While some provinces and territories disaggregate performance for a wide range of variables, others only disaggregate according to a small number, or do not report performance at all. Ontario, followed by British Columbia, was the province with the most extensive disaggregation of performance results (see Table 2). Saskatchewan,[2] Quebec, Prince Edward Island, and Nunavut did not publicly report performance. While most provinces and territories reported disaggregated data in some respect, rarely were the implications of the data discussed, especially in terms of validity. For example, although the British Columbia Ministry of Education (2013) provided a 46-page document of tables and graphs reporting aboriginal education statistics, it did not discuss the implications or significance of these results. Of all the provinces and territories, Ontario most extensively discussed disaggregated performance, publishing articles such as "The Unnecessary Lag in Boys' Achievement" (Education Quality and Accountability Office, 2010).

Construct Remodeling

The process of disaggregating performance provides information about differences in populations' test scores; it does not, however, on its own, explain what the reasons for these differences are. Differences in scores could be a reflection either of actual differences in ability between the groups or of differences in how the tests functioned across these different populations. Examining response processes of test-takers can be an effective way of helping test-users understand the reasons for these differences in population scores.

Fox and Cheng (2007) studied the experiences of 22 students for whom English was a first language (L1) and 136 students for whom English was a second language (L2) who were taking a sample version of the Ontario Secondary School Literacy Test to determine the degree to which the OSSLT was measuring the same construct for different populations of students. The study found that the test was measuring different constructs for L1 and L2 students and that construct irrelevant variance was a primary reason for this problem. Prompts for writing tasks, they found, relied on single words or phrases that were devoid of contextual information that would help students understand what they were being asked to write. In many instances, L2 students who did not initially write a response to these questions were able to

Table 2. Disaggregation of Large-Scale Writing Assessment Results by Province

Category	British Columbia	Alberta	Saskatchewan	Manitoba	Ontario	Quebec	New Brunswick	Nova Scotia	Newfoundland and Labrador	Prince Edward Island	Northwest Territories	Nunavut	Yukon
District/Region	✔				✔		✔	✔	✔		✔		
School					✔			✔	✔				
Cohort (test year)	✔	✔		✔	✔		✔	✔	✔		✔		✔
Academic Level					✔			✔	✔				
Gender	✔				✔		✔		✔				✔
Aboriginal Status	✔												✔
ELL Status	✔				✔								
Language Stream				✔	✔		✔						
Special Needs Status	✔				✔								
Mobility					✔								
Eligibility					✔								

successfully complete these assignments when the words were explained to them. Based on these data, Fox and Cheng claim that the OSSLT was measuring a language proficiency construct rather than a writing construct for L2 students. They also found that L2 students often did not fully understand the large-scale test genre in the same way that their L1 counterparts did, and that consequently for these L2 students, the test was measuring knowledge of the testing genre in addition to the writing construct it was intended to measure.

Fox and Cheng (2007) claim that this construct issue "may be contributing to an unfair advantage or disadvantage for some test-takers" (p. 22). Flowing from this, they note that "there are important consequences of OSSLT failure as it may lower confidence and self-esteem and increase perceptions of difficulty of some L2 test-takers" (p. 22). This finding is particularly troubling in light of the real possibility that these negative affective consequences are a result of problems with the test's construct validity rather than with the students' own abilities.

Implications: Intended

Because large-scale testing programs are designed to achieve specific purposes, it is important to evaluate the degree to which these programs achieve their purposes. In the Canadian context, many provinces do not explicitly identify school improve-

ment as a goal of their testing programs; rather, this goal is couched in terms of monitoring educational systems and providing stakeholders with information they can act upon to improve performance. In the documentation these testing agencies provide to the public, they trumpet the use of system-wide data to help educators improve student achievement. The EQAO (2011), for example, claims that "since results [became] available for every student, provincial test data have become a key ingredient in helping schools, school boards and the province identify students' strengths and target areas where attention and resources are needed" (p. 2). In support of this claim, they assert that 96% of school principals in Ontario use EQAO data to guide school improvement initiatives. The EQAO's claims are supported by Hardy (2010), as well as Anderson and Macri (2009), who found that the district leaders and principals they interviewed used large-scale test data to identify schools, curricular areas, and students who were in need of greater support. The interviewees reported using these data to "develop school improvement plans and individual student interventions" (p. 203). Similarly, Hardy's (2010) participants reported an increased focus on literacy education and literacy-oriented professional development as a result of the institution of the OSSLT. They also saw EQAO data as supporting a data-driven approach to school improvement.

In reporting on the achievement of program outcomes, test users and designers need to take a critical approach, however. For example, they need to recognize that principals' use of test data for school improvement purposes does not automatically translate into actual improvements in education. Anderson and Macri (2009) problematize the EQAO's reporting on principals' use of assessment data, observing that these same principals resisted the narrow focus of the Ministry's accountability system in favor of a broader focus on the development of the whole child. Similarly, Hardy (2010) observes that even though principals pushed for improved EQAO scores, "there is also evidence of concern among principals about the negative effects of a strong emphasis upon generic conceptions of literacy" (p. 432) measured by these tests. Consequently, these principals were conflicted in their work, feeling a need to be "acquiescent to more managerial pressures for more restricted, standardized conceptions of learning, but also strongly focused upon the needs of specific students and specific schools" (p. 433).

The concerns raised by these administrators reflect the importance of an integrated approach to validation. Claims that intended outcomes are being met are bolstered when other sources of validity evidence support them, and are undermined when they do not. The concerns raised by Hardy's participants reflect the construct validity issues raised by many of the other studies concerning the OSSLT that we have discussed in this article, especially in regard to the issue of construct underrepresentation; further, their concerns about the narrowing of educational focus are borne out in the studies that directly examined teacher and student experiences.

Implications: Unintended

The study of unintended outcomes or implications is equally dependent on the integration of validity evidence. An important starting point for an analysis of unintended outcomes is construct validity evidence. If construct validity data point

to issues of underrepresentation or irrelevant variance, then a close examination of what consequences these construct flaws carry for individuals, populations, and broader educational systems is necessary. It is important to recognize that tests are never neutral. In many instances, large-scale, government-mandated tests represent the most concrete statements about what knowledge and skills are being valued within an educational system, and as such, they carry significant potential to (mis)shape student and teacher values within that system.

Many of the studies we reviewed reported negative unintended consequences at the systems level linked to construct validity issues. Skerrett and Hargreaves (2008) noted that the OSSLT supported traditional rather than innovative approaches to literacy education. They observed that "standardization reinforced and validated the traditional curriculum and teaching strategies of veteran teachers who lacked professional training or experience with diversity" (p. 935). At the same time, Skerrett (2010) observed that the OSSLT worked against innovation in education by stunting the growth of professional learning cultures in the schools she studied and by compelling innovative teachers to compromise their professional judgment in the following ways: teachers focused on teaching testable literacy skills even though they felt that the test did not accurately measure the knowledge and skills they believed to be most important; and teachers became overreliant on standardized rubrics, leading them to narrow their assessment practice, which in turn limited student opportunities to demonstrate learning across multiple contexts and modalities. Ricci's (2004) case study generated similar findings. He reports that in the school he studied, class time was diverted from regular instruction to focus on test preparation (one period per week, plus a solid five- to six-week unit focusing on the test); preparation time focused on low-level skill-and-drill work rather than higher-order literacy skills; teachers fell substantially behind in their course material; and teachers felt compelled to prepare students for the test even though they questioned its usefulness and validity. These findings reinforce administrators' concerns, cited by Hardy (2010) and Anderson and Macri (2009), that the narrow focus of the OSSLT diverts school attention away from higher-order literacy skills in favor of traditional literacies emphasized on the test. Collectively, this body of research raises questions about the value of greater emphasis on literacy education in the context of these tests if that also means an increased focus on teaching to a narrow and limited construct.

Problems with respect to the impact of these tests on different populations of students are also discussed extensively in the research literature (see our earlier discussion on construct remodeling). Studies have found, for example, that the OSSLT's scores reflect different constructs for different populations of students and that these differences may result in unwarranted negative consequences for some students (Fox & Cheng, 2007; Kim & Jang, 2009).

Summary and Significance

By developing a framework for collecting consequential validity evidence, we were able to integrate previously unconnected consequential validity studies to demon-

strate the challenges facing large-scale, government-mandated Canadian writing assessment programs as they seek to measure performance on a complex construct across a diverse population of students and contexts. Each study examines one aspect of the validity question largely independent of the others, thus providing a limited picture of the consequential validity issues associated with the tests the study is evaluating. Using our framework to examine the studies as instrumental cases, however, we see a much stronger and clearer picture of consequential validity issues at play and the ways in which diversity is restricted and constrained within these assessment contexts.

To illustrate the integrated validity argument that can be developed when the diverse sources of evidence discussed in this paper are combined, we have organized the concerns raised by these studies according to three general themes.

First, the tests constrained writing as a construct for the following reasons:

- ignoring emergent forms of digital writing and multiliteracies (Lotherington, 2004; Peterson et al., 2012);
- truncating the writing process, emphasizing only the final draft while ignoring process-oriented writing (Peterson et al., 2012; Slomp, 2005, 2008);
- focusing more on procedure (write *X* number of lines) than actual writing (Ricci, 2004);
- focusing on narrow/traditional constructs of writing (Hardy, 2010; Peterson & McClay, 2010; Slomp, 2005, 2008); and,
- ignoring the importance of multiple assessments over time, as well as differentiated forms of assessment (Peterson & McClay, 2010; Skerrett & Hargreaves, 2008).

Second, rather than encouraging a broadening of instruction, these tests were found to limit pedagogical diversity in the classrooms studied for the following reasons:

- failing to provide teachers with any new evidence about their students (Lam & Bordignon, 2001; Skwarchuk, 2004; Toohey, 2007);
- taking time away from teaching to focus on test-taking skills (Lam & Bordignon, 2001; Ricci, 2004; Skwarchuk, 2004; Slomp, 2005, 2008); and
- encouraging convergent thinking over divergent (creative) thinking (Ricci, 2004; Zheng, Klinger, Cheng, Fox, & Doe, 2011).

Third, the studies suggest that these testing programs undermined diversity through their negative impacts, especially on marginalized populations of students and their teachers (Cheng, Klinger, & Zheng, 2009; Doe, Cheng, Fox, Klinger, & Zheng, 2011; Fox & Cheng, 2007; Kim & Jang, 2009; Skwarchuk, 2004; Toohey, 2007). The studies reviewed found that the tests have the following deleterious effects:

- lacked cultural diversity and favored students who had knowledge of Canadian symbols, historical events, and artifacts (Doe et al., 2011; Kim & Jang, 2009);

- led to students' negative self-image as writers and lessened motivation to learn (Doe et al., 2011; Fox & Cheng, 2007; Kearns, 2011); and
- disempowered teachers while undermining their professional judgment (Skerrett, 2010; Skerrett & Hargreaves, 2008).

Collectively, our findings challenge the idea that these testing programs are supporting improvements in writing education in Canada, especially with respect to fostering and supporting diversity within the system. These findings suggest the need for a reexamination of the writing assessment programs currently in use in Canada.

Limitations

We present the above summary of findings as an illustration of the integrated argument that can be made when drawing on our model for collecting consequential validity evidence. Basing an analysis of research and government reports on a model that was not used to guide those original studies, we recognize, is problematic. Publicly available government reports, while an important source of data, may not for political reasons present a full accounting of consequential validity data collected with respect to these testing programs. Utilizing freedom-of-information legislation to gain access to nonpublicized information could potentially have provided a fuller accounting of these data. Additionally, the research we reviewed for this study originated in assessment, literacy, and school leadership research traditions. A number of these studies, while implicitly dealing with validity issues, were not informed by validity theory. Consequently, the quality of this research with respect to how it engaged with the issue of validity was uneven. Finally, because the studies we report on as instrumental cases were not conceptualized within a single cogent research program, we present our summary of this evidence with some degree of tentativeness, recognizing that this integration of studies leaves us open to the critique of having compared apples to oranges. While we do find the degree to which these studies agree with one another to be compelling in its own way, we reiterate that our primary intention in this article is to explicate and illustrate a model for collecting consequential validity evidence within a new framework, not to make a definitive statement about the state of writing assessment in Canada.

Directions for Further Research

Through our development and use of a framework for collecting consequential validity evidence, our study reveals important avenues for continued research.

First, the studies we reviewed revealed that the stakeholders farthest from the classroom viewed large-scale assessment most positively. Educational administrators noted that large-scale assessments provided a data-driven approach to school and district improvement, while also providing a focus for professional development (Anderson & Macri, 2009; Hardy, 2010). Teachers and students, conversely, mostly held negative attitudes toward large-scale testing. Thus, future studies might reveal how the needs of multiple stakeholders might be better balanced.

Second, our study found gaps in the collection of consequential validity evidence for Canadian writing assessment programs. We did not find studies on consequential validity issues regarding design processes, sampling plan designs, or data disaggregation. These gaps should be filled.

Third, while the impact of these assessment programs on ELLs was an important element of this corpus of studies, we found no studies that examined these issues with respect to the Aboriginal students in Canadian schools, suggesting that a systematic examination of the impact of large-scale writing assessment on Aboriginal students in Canada is needed.

Fourth, our review of the literature also found that the consequential validity research in Canada was uneven in quality with respect to its design and execution. Framing future research around consequences within a defined validation model such as the one used in the present research will lend greater consistency to the conceptual development of these studies without, we believe, limiting the range of possible research designs. None of the studies we reviewed for this paper would have required methodological changes to fit within the validity framework we have articulated in this paper.

Fifth, for too long writing studies scholars and educational measurement researchers have been engaged in parallel discourse on the work of large-scale writing assessment (Huot, 2002). Rarely do these avenues of discourse meet, but this needs to change. Adler-Kassner and O'Neill (2010) suggest that writing assessment scholars "must understand the dominant frame (and its attendant values and ideologies) and work with it—even within it, if necessary" to ensure that these assessments reflect the theories and values of our discipline (p. 177). Framing future work on the consequences of large-scale writing assessment practices under the umbrella of validity theory will enable writing assessment researchers to better find and build upon each other's work.

Finally, a fundamental flaw in many test-based accountability programs is that they beg the question of consequences; that is, they assume that testing programs will improve teaching and learning, but they do not collect direct evidence to investigate such assumptions (Kane, 2006). Our findings confirm this flaw. We believe that this gap in research exists in part because a systematic approach to collecting and analyzing consequential validity research has not previously been articulated. Having articulated such a model, we believe that a comprehensive program of comparative studies across multiple provincial, territorial, and even national contexts would reveal a great deal about the consequential validity issues at play in the design and implementation of large-scale, government-mandated writing assessments. Such studies, we believe, will provide a far more powerful body of evidence to inform ongoing debates about the value of large-scale assessment programs.

Directions for Assessment Policy

The expectation that assessment developers and users should collect and report consequential validity evidence is not new; it has been part of the discussion on validity for more than 60 years now. During that time, the assessment community

has been engaged in a series of nuanced debates, both about whose responsibility it is to collect these data and about the extent of the consequences that test users and developers should be responsible for. In the meantime, little consequential validity research has been conducted or reported by the testing agencies or the governments that contract them. Assessment stakeholders seem to embrace the idea of using tests to hold others accountable, but fail to systematically address the one aspect of validity theory that turns that accountability mechanism back on ourselves: the implications of our assessment actions. Brennan (2006) suggests that the agencies mandated to hold students and teachers to account resist research into the effects of their programs because they are concerned about what this research might find. He writes:

> Publicly available, timely documentation related to validity arguments is often the exception rather than the rule. The uncomfortable reality is that if such documentation is clear, complete, and forthright, it will not always support validity arguments. (p. 8)

Brennan's suggestion highlights, most clearly of all, the need for research into the consequences of large-scale assessments for systems of education. If the true goal of government-mandated, large-scale assessment programs is to improve systems of education, then it is only logical that the designers and users of these tests should examine and publicly report on the consequences that accrue as a result of the use of these tests. We hope that the framework presented in this paper will enable the collection and reporting of these important data.

ACKNOWLEDGMENTS

This study was supported by a grant from the Social Sciences and Humanities Research Council of Canada. The authors would like to express their deep gratitude to Norbert Elliot for his extensive feedback and support over several months during the revision of this article. As a reviewer of an early draft of this article, he understood what we were trying to accomplish and provided us with the critical insight necessary to bring these revisions to fruition. His generosity of spirit and insight are greatly appreciated. We would also like to thank Jane O'Dea, John Rymer, and Bob Broad and the students in his graduate course on writing assessment for their feedback on early drafts of this work. Finally, we would also like to thank Mya Poe for her exceptional work as guest editor of this special issue.

NOTES

1. For an example of a more dialogic approach to writing assessment design, see Broad's (2003) explication of Dynamic Criteria Mapping.
2. Although the province does not report the results of large-scale assessments, the majority of school districts share their Continuous Improvement Reports online on their division websites each year.

REFERENCES

Adler-Kassner, L., & O'Neill, P. (2010). *Reframing writing assessment to improve teaching and learning.* Logan: Utah State University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests.* Washington, DC: American Educational Research Association.

Anderson, S. E., & Macri, J. R. (2009). District administrator perspectives on student learning in an era of standards and accountability: A collective frame analysis. *Canadian Journal of Education, 32*(2), 192–221. Retrieved from http://search.proquest.com/docview/215372058?accountid=12063.

Beaufort, A. (2007). *College writing and beyond: A new framework for university writing instruction* (pp. 1–242). Logan, UT: Utah State University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger.

British Columbia Ministry of Education. (2013). *Aboriginal report 2007/08–2011/12: How are we doing?* Retrieved from http://www.bced.gov.bc.ca/reports/pdfs/ab_hawd/Public.pdf.

British Columbia Ministry of Education. (n.d.). *Grade 10 Provincial Examination Specifications.* Retrieved from http://www.bced.gov.bc.ca/exams/specs/grade10/en/2011.htm.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*, 213–260.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing.* Logan: Utah State University Press.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97–113.

Camp, H. (2012). The psychology of writing development—And its implications for assessment. *Assessing Writing, 17*, 92–105.

Cheng, L., Klinger, D. A., & Zheng, Y. (2009). Examining students' after-school literacy activities and their literacy performance on the Ontario Secondary School Literacy Test. *Canadian Journal of Education, 32*(1), 118–148.

Citizenship and Immigration Canada. (n.d.). *Canadian multiculturalism: An inclusive citizenship.* Retrieved from http://www.cic.gc.ca/english/multiculturalism/citizenship.asp.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*(5), 732–743.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Denzin, N. K. (2010). Moments, mixed methods, and paradigm dialogs. *Qualitative Inquiry, 16*(6), 419–427.

Doe, C., Cheng, L., Fox, J., Klinger, D., & Zheng, Y. (2011). What has experience got to do with it? An exploration of L1 and L2 test takers' perceptions of test performance and alignment to classroom literacy activities. *Canadian Journal of Education, 34*(3),

68–85. Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=EJ946084&ERICExtSearch_SearchType_0=no&accno=EJ946084.

Education Quality and Accountability Office. (2010). *The unnecessary lag in boys' achievement.* Retrieved from http://www.eqao.com/Publications/ArticleReader.aspx?Lang=E&article=b10A001&section=.

Education Quality and Accountability Office. (2011). *Information bulletin: Ontario Secondary School Literacy Test (OSSLT), 2010–2011.* Retrieved from http://www.eqao.com/pdf_e/11/Cib_Xe_0611_web.pdf.

Education Quality and Accountability Office. (2012). *EQAO's provincial secondary school report.* Retrieved from http://www.eqao.com/ProvincialReport/Files/12/PDF/EQAO_PR_10e_FTE_Achievement Results_0912_Web.pdf.

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis, J. Burstein, & S. Apel (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 16–35). London: Routledge.

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education, 14*(1), 9–26.

Gould, S. J. (1996). *The mismeasure of man.* New York: Norton.

Hardy, I. J. (2010). Leading learning: Theorizing principals' support for teacher PD in Ontario. *International Journal of Leadership in Education, 13*(4), 421–436.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning.* New York: Teachers College Press.

Huot, B. (2002). *(Re)articulating writing assessment for teaching and learning.* Logan: Utah State University Press.

Inoue, A. B. (2009). The technology of writing assessment and racial validity. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education.* Hershey, PA: Information Science Reference.

Kane, M. T. (2006). Validation. In Robert L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–66). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kearns, L. (2011). High-stakes standardized testing and marginalized youth: An examination of the impact on those who fail. *Canadian Journal of Education, 34*(2), 112–130. Retrieved from http://ojs.vre.upei.ca/index.php/cje-rce/article/view/354.

Kim, Y., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning, 59*(4), 825–865.

Klinger, D. A., Rogers, W. T., Anderson, J. O., Poth, C., Calman, R. (2006). Contextual and school factors associated with achievement on a high-stakes examination. *Canadian Journal of Education, 29*(3), 771–797. Retrieved from http://www.csse-scee.ca/CJE/Articles/FullText/CJE29-3/CJE29-3-Klingeretal.pdf.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: Praeger.

Lam, T. C., & Bordignon, C. (2001). An examination of English teachers' opinions about the Ontario grade 9 reading and writing test. *Interchange, 32*(2), 131–145.

Lane, S. (2013). The need for a principled approach for examining indirect effects of test use. *Measurement: Interdisciplinary Research and Perspectives, 11*(1–2), 44–46.

Lotherington, H. (2004). Emergent metaliteracies: What the Xbox has to offer the EQAO. *Linguistics and Education, 14*, 305–319.

Lynne, P. (2004). *Coming to terms: A theory*

of writing assessment. Logan: Utah State University Press.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research, 40*, 109–126.

McLuhan, M. (1962). *The Gutenberg Galaxy: The making of typographic man.* Toronto, Canada: University of Toronto Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5–12.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*, 3–9.

Nova Scotia Department of Education. (2010). *Minister's report to parents and guardians: 2010 student assessment results.* Retrieved from http://mrpg.ednet.ns.ca/sites/default/files/mrpg-all.pdf

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.

Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 107–124). Washington, DC: SAGE.

Peterson, S. S., & McClay, J. (2010). Assessing and providing feedback for student writing in Canadian classrooms. *Assessing Writing, 15,* 86–99.

Peterson, S. S., McClay, J., & Main, K. (2012). An analysis of large-scale writing assessments in Canada (grades 5–8). *Alberta Journal of Educational Research, 57,* 424–445. Retrieved from http://ajer.synergiesprairies.ca/ajer/index.php/ajer/article/view/947.

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*, 9–13.

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice, 18*, 13–17.

Ricci, C. (2004). The case against standardized testing and the call for a revitalization of democracy. *Review of Education, Pedagogy, and Cultural Studies, 26*(4), 339–361.

Roos, N. P., Brownell, M., Guevremont, A., Fransoo, R., Levin, B., MacWilliam, L., & Roos, L. L. (2006). The complete story: A population-based perspective on school performance and educational testing. *Canadian Journal of Education, 29*(3), 684–705. Retrieved from http://files.eric.ed.gov/fulltext/EJ756118.pdf.

Schönemann, P. H. (1997). On models and muddles of heritability. *Genetica, 99*(2–3), 97–108.

Skerrett, A. (2010). "There's going to be community. There's going to be knowledge": Designs for learning in a standardised age. *Teaching and Teacher Education, 26*, 648–655.

Skerrett, A., & Hargreaves, A. (2008). Student diversity and secondary school change in a context of increasingly standardized reform. *American Education Research Journal, 45*(4), 913–945.

Skwarchuk, S. (2004). Teachers' attitudes toward government-mandated provincial testing in Manitoba. *Alberta Journal of Educational Research, 50,* 252–282. Retrieved from http://ajer.synergiesprairies.ca/ajer/index.php/ajer/article/view/486.

Slomp, D. H. (2005). Teaching and assessing language skills: Defining the knowledge that matters. *English Teaching: Practice and Critique, 4*(3), 141–155. Retrieved from http://education.waikato.ac.nz/research/files/etpc/2005v4n3art8.pdf.

Slomp, D. H. (2008). Harming not helping: The impact of a Canadian standardized writ-

ing assessment on curriculum and pedagogy. *Assessing Writing, 13,* 180–200.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. R. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3–21). New York: Routledge.

Stake, R. (1995). *The art of case study research.* Newbury Park, CA: SAGE.

Stake, R. (2005). Qualitative case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (3rd ed., pp. 443–466). Thousand Oaks, CA: SAGE.

Statistics Canada. (2011a). *2011 national household survey: Aboriginal peoples in Canada: First Nations people, Métis and Inuit.* Retrieved from http://www.statcan.gc.ca/daily-quotidien/130508/dq130508a-eng.htm.

Statistics Canada. (2011b). *2011 census of population: Linguistic characteristics of Canadians.* Retrieved from http://www.statcan.gc.ca/daily-quotidien/121024/dq121024a-eng.htm.

Statistics Canada. (2011c). *Immigrant and ethnocultural diversity in Canada.* Retrieved from http://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-010-x/99-010-x2011001-eng.cfm.

Toohey, K. (2007). Are the lights coming on? How can we tell? English language learners and literacy assessment. *Canadian Modern Language Review, 64*(2), 253–272.

White, E. M., Elliot, N., & Peckham, I. (in press). *Very like a whale: The assessment of writing programs.* Logan: Utah State University Press.

Yancey, K. (2009). 2008 NCTE presidential address: The impulse to compose in the age of composition. *Research in the Teaching of English, 43,* 316–338.

Zheng, Y., Klinger, D. A., Cheng, L., Fox, J., & Doe, C. (2011). Test-takers' background, literacy activities, and views of the Ontario Secondary School Literacy Test. *Alberta Journal of Educational Research, 57,* 115–136. Retrieved from http://ajer.synergiesprairies.ca/ajer/index.php/ajer/article/view/900.

David H. Slomp is an assistant professor of literacy assessment at the University of Lethbridge, where he co-directs the Masters in Education in Curriculum and Assessment program. His research into the consequential validity of large-scale writing assessment in Canada is funded through a grant from the Social Sciences and Humanities Research Council of Canada. His work has appeared in *College Composition and Communication* and in *Assessing Writing.*

Julie A. Corrigan is a Bombardier Canada Graduate Scholar and PhD candidate at the University of Ottawa, Canada. Last year she attended the University of Connecticut's New Literacies Research Lab as a visiting scholar. Her doctoral work will continue to investigate online writing assessment.

Tamiko Sugimoto is a master of education student at the University of Lethbridge. She obtained her bachelor of arts (with distinction) in psychology from the University of Lethbridge in 2005. She is currently completing her thesis investigating the cessation of nonsuicidal self-injury.

Initial submission: March 2, 2013
Final revision submitted: September 5, 2013
Accepted: September 25, 2013