An Integrated Design and Appraisal Framework for Ethical Writing Assessment

by David Slomp, University of Lethbridge

In my introduction to this special issue, I highlighted the insufficiency of key measurement concepts--fairness, validity, and reliability--in guiding the design and implementation of writing assessments. I proposed that the concept of ethics provides a more complete framework for guiding assessment design and use. This article advances the philosophical foundation for our theory of ethics articulated by Elliot (this issue). Starting with fairness as first principle, it examines how safety and risk can be addressed through the application of an integrated design and appraisal framework (IDAF) for writing assessment tools. The paper is structured around two case studies set in Alberta, Canada. Case Study 1 applies Kane's (2013) IUA model of validation to an appraisal--Alberta's English 30-1 (grade 12 academic English) diploma exam program--highlighting in the process the limitations in contemporary validity theory. Case Study 2 examines an assessment design project I am currently undertaking in partnership with 8 English language arts teachers in southern Alberta. This case study examines how the IDAF supports ethical assessment design and appraisal.

1.0 Background

Historically, the measurement community has claimed that validity must be the primary concern in the design and appraisal of assessment programs. In this SI, we collectively argue against this understanding, suggesting instead that ethics must supplant validity in this respect. To make clear why this change is necessary I have developed this article as two case studies, both set in Alberta, Canada.[1] The first case study (§3, §4, and §5) provides a brief illustrative application of Kane's (2013) IUA model of validation to Alberta's English 30-1 diploma exam program. While there are other approaches to validation available to consider, I have chosen to focus on Kane's model because it is the most advanced. This case study will demonstrate that in spite of its many strengths, Kane's IUA model is limited by its focus on interpretation and score use, and by its siloed approach to examining issues of fairness, reliability, and validity. The second case study (§6) examines an assessment design project I am currently conducting with eight junior and senior high school teachers in southern Alberta. In this case study I present and apply an integrated design and appraisal framework (IDAF) to the design of classroom-based writing assessments. This case study demonstrates how the IDAF compels attention to the six dimensions of the theory of ethics elucidated within this special issue: sociocultural perspectives, access, opportunity to learn, maximum construct representation, disaggregation of data, and justice.

Each of these dimensions is important to the design and appraisal of high quality assessments. A focus on *sociocultural perspectives* enables test designers to recognize that the constructs upon which we base our assessments are socially constructed and potentially unstable across social, cultural, and temporal contexts. A focus on *access* and *opportunity to learn* highlights the importance of ensuring all test takers should be provided with equal opportunity to demonstrate their standing with respect to the construct of interest—this includes equity with respect to the educational opportunities students receive, and equity with respect to how test items function for diverse populations of students. Attention to *maximum construct representation* is intended to ensure assessment programs promote teaching and learning that is focused on our best understandings of the constructs of interest. *Disaggregation of data* and *justice* focus attention on the effect of assessment programs on populations of students, and emphasize the goal that unwarranted negative consequences stemming from the design and implementation of an assessment program are identified and remedied. By attending to each of these dimensions of our theory of ethics, this IDAF is designed to assist in the development of high-quality writing assessments that better serve students, teachers, and educational systems.

2.0 Validity Theory: Its Promise and Its Problems

Validity theory has much to offer with respect to guiding ethical practice. Cronbach (1970) observed that a concern for social consequences and ethics already had been linked to the core function of assessment in the 19th century when reformers like Galton, Jefferson, and Mill saw testing as a tool for creating an open system of competition, one that would enable society to apportion opportunities to people based on aptitude rather than heritage. The purpose of criterion-reference validation that emerged by the 1930s was to examine how well assessment instruments were delivering on this intended social consequence. To this end, the *Standards for educational and psychological tests and manuals* (AERA, APA, NCME, 1966) explicitly linked validity and social consequences:

> Almost any test can be useful for some functions and in some situations, but even the best test can have damaging consequences if used inappropriately…. It is not inconceivable that many test makers and test publishers could fulfill to a highly adequate degree most of the standards presented and still produce a test that would be virtually worthless from the standpoint of fulfilling its intended or stated objectives. (p. 6)

This caution is important. Already in 1966 the *Standards* recognized their own limitations in protecting society from "damaging consequences" of even the most conscientiously designed assessments.

As validity theory evolved, this concern for social consequences remained. In the 1980s Messick explicitly fused the concern for

construct validity and social consequences into an integrated evaluation of an assessment's use. He claimed negative social consequences undermine a validity argument if those consequences are shown to be a product of either construct underrepresentation or construct irrelevant variance (Messick, 1989). The significance of Messick's argument was that it obligated assessment developers to pay attention both to how well their assessment was capturing the constructs it was designed to measure and to the social consequences that accrue as a result of its development and use.

A limitation in Messick's work, however, was that it failed to provide necessary guidance in how to conduct validation studies that reflected this unified understanding. As a result, the field has focused extensively on the question of agency, parsing out who (test developers or test users) was most responsible for attending to the issue of consequences (Nichols & Williams, 2009). While the debate rolls on, little validation research exploring the social consequences of assessment has been conducted (Cizek, Bowen, & Church, 2010; Lane, 2013; Zumbo & Chan, 2014). And so the scenario in which assessments could be designed in compliance with the *Standards* but with limited concern for consequences (either intended or unintended) has clearly materialized (Brennan, 2006).

1. *Scoring Inference:* An inference that leads from observed performance to an observed score.
2. *Generalization inference:* Extrapolates from observed scores to a larger universe of possible performances on similar measures.
3. *Extrapolation inference:* Extends the interpretation beyond the specific test context, construct, or set of tasks to the broader universe of possible performances.
4. *Decision Inference:* Takes us from a person's score to a decision based on that score.

Kane stressed that each inference is essential; if one inference fails, the entire argument fails. Similarly, weakness in one inference cannot be compensated for by strength of the other inferences.

While helpful, both Messick's and Kane's frameworks are limited by their narrow attention to score use and interpretation. These frameworks are effective in helping to address a range of social consequences pertaining to a test's effect on students while offering little in the way of addressing issues related to a test's effect on teachers and educational systems. To illustrate these strengths and weaknesses, I explore the validity argument for Alberta's English 30-1 (grade 12 academic English) writing exam.

In §3 of this paper I provide a description of the English 30-1 diploma exam. In §4 I examine in order each inference needed for a robust a program of validation for this exam. Finally, in §5 I discuss how well this validity argument addresses the six dimensions of our theory of ethics.

3.0 Alberta's English 30-1 Diploma Exam

Alberta's English 30-1 diploma exam program has been in use since 1984. Over the years, those responsible for this exam have continually asserted its international reputation in defense of its continued use (Staples, 2012). While there isn't space in this article to do justice to a full evaluation of the validity argument for this exam, it is worth looking at a few key elements of such an argument, to illustrate the limitations of validity theory.

The exam is in two parts. Part 1 is a timed, impromptu writing exam; Part 2 is a reading comprehension exam. Students are given three hours to complete each part. Part 1 (the focus of this discussion) consists of two written responses: A personal/creative response, and a critical analytical response. Both constructed response (CR) tasks (Bennett, 1993) are thematically linked to each other.

The purpose of this exam program is to certify the level of individual student achievement in grade 12 English language arts; to ensure province-wide standards of achievement are maintained; and to report individual and group results (Alberta Education, 2004).

4.0 A Program of Validation for Alberta's English 30-1 Diploma Exam

A strong program of validation research would flow from an analysis of the scoring inference, to the generalization inference, the extrapolation inference, and decision inferences (Kane, 2006, 2013).

4.1 The Scoring Inference

The scoring inference for Alberta English 30-1 writing exam would claim:

The English 30-1 diploma exam's scoring procedures are appropriate, are implemented correctly, consistently, and free from overt bias.

4.2 Generalization Inference

The next step in this process would be to test the generalization inference, which in this case might claim:

The writing component of the English 30-1 writing exam has been designed so tasks, contexts, occasions, and raters collectively provide for a consistent measurement of test-taker ability with respect to the construct of "writing ability."

The English 30-1 diploma exam was designed for consistency and as such is administered in a controlled, standardized environment monitored by certified teachers. The challenge when measuring complex constructs such as writing ability, however, is that the very design choices aimed at achieving consistency work against the second component of the generalization inference. In such situations, test developers often narrow their measurement to a representative sample of the target domain (Kane, 2013). Problems occur when that representative sample ignores important aspects of the construct being measured. For example, Alberta Education's decision to use a timed impromptu exam format instead of a portfolio-based assessment introduces systematic error into the exam context. Students who write using an elaborated recursive process can be systematically disadvantaged by a writing exam that compels them to truncate this process (a key element of the writing construct). For these students, it is difficult, if not impossible, to generalize from their performance on this exam to their performance on other non-test writing tasks, especially if those tasks allow for a more elaborated process.

4.3 Extrapolation Inference

The extrapolation inference expands the focus of the validity argument from the construct sample to the broader target domain. In this case, the core claim might be:

The English 30-1 test scores accurately reflect student's writing ability with respect to the full range of writing tasks in the target domain, including nontest performances in nontest contexts.

This inference relies on an analysis of both the target domain measured by the test and of the broad range of construct-related tasks found within the universe of possible tasks envisioned by the exam's purpose.

A key purpose of the exam is to certify level of achievement in English 30-1. The universe of generalization appropriate to this exam, then, includes the range of writing tasks and skills envisioned by the English 30-1 program of studies.

The Alberta curriculum envisions students who are able to deconstruct rhetorical situations with a view to selecting strategies and approaches that will help them effectively write for that situation. It projects a vision of a student writer who is flexible in his or her thinking, able to adapt, ready to reconsider first ideas, and able to negotiate a complex recursive writing process. Additionally, the curriculum expects students to develop the capacity to generate prose, poetry, creative fiction, non-fiction, and multimodal texts (Slomp, 2008, pp. 182-183).

Even under the most cursory of reflections, it is painfully obvious that no single-sitting, timed, impromptu writing task, no matter how well constructed, could ever measure such a broad range of complex learning outcomes and their associated constructs. This is especially true if the CR task is essentially the same for each iteration of the exam (as is the case with the English 30-1 writing exam). Attention to the scoring inference has undermined the strength of the extrapolation inference.

4.4 Decision Inference

The final stage in this validation process is to investigate the decision inference, the most complex of the four types of inferences because it depends on evidence from each of the previous inferences.

The decision inference for the English 30-1 diploma exam might read as follows:

Decisions to award high school diplomas based in part on student's English 30-1 exam test scores are justified. In using the exam scores for this purpose intended outcomes are being achieved at a reasonable cost and with limited negative consequences.

The evidence required to support this inference is varied, and depends on specified use.

Part one of testing the decision inference requires an assessment of the decision to use the exam score for the purpose of awarding a high school diploma. Justifying this decision requires a content analysis of the diploma exam (Kane, 2002). The *Information Bulletin* (Alberta Government, 2015) for the English 30-1 exam provides little evidence to support this aspect of the decision inference. The document demonstrates that each iteration of the exam only partially measures 68% of the English 30-1 curriculum outcomes. Alberta Education's own documentation demonstrates that the exam consistently underrepresents student performance

on curriculum when compared with teacher awarded marks (Alberta Government, 2015).

Testing the second half of the decision inference involves a cost-benefits analysis regarding the use of the test itself. This analysis focuses on weighing positive or negative, intended and unintended consequences against one another. It should examine intended outcomes, adverse impacts, and systemic effects (Kane, 2013).

4.4.1 Intended outcomes. Alberta's English 30-1 diploma exam program has three intended purposes: to report individual and group results; to ensure province-wide standards of achievement are maintained; and to certify the level of individual student achievement in grade 12 English language arts (Alberta Education, 2004).

These purposes are achieved to some degree: Each year Alberta Education releases reports both to individual students and to school districts detailing student performance on the English 30-1 diploma exam. Alberta Education has adopted a policy of equating test scores. Performance across test iterations is analyzed to determine whether or not performance levels are increasing or decreasing.

4.4.3 Systemic effects. Analysis of systemic effects begins by acknowledging that testing programs are designed to help students in some way and then proceeds by examining the extent to which such benefits are realized (Kane, 2013). Kane limited this analysis to intended populations (in this case concern for students but not for teachers). Addressing the consequences for intended population(s) requires examining how the reporting of assessment results, the maintenance of standards, and the certification of achievement benefit students as a whole. Alberta Education claims their exam program promotes confidence in the system, supports fair interpretation and use of test scores, and enhances transparency and comparability of school awarded marks. Within the domain of public perception, this claim appears to be warranted; public support for Alberta's assessment programs is high and widespread (Webber, Aitken, Lupart, & Scott, 2009).

My own research into the English 30-1 diploma exam, however, found that in the context of this assessment teachers narrowed their marking criteria, the number of genres they assigned, and the range of curriculum outcomes they assessed to the narrow set of criteria, tasks, and outcomes measured by the diploma exam (Slomp, 2008; Slomp, Graves, & Broad, 2014). In spite of this narrowing of instruction, Alberta Education reports that Alberta's English 30-1 diploma exam systematically and consistently underrepresents student performance in English 30-1 when compared to teacher awarded marks, and that it does this at twice the rate for the highest achieving students (Alberta Government, 2015).

As a further consequence, students learned to value the narrow set of skills measured by the diploma exam in their own development as writers (Slomp, 2007) and teachers report that given the stakes, students are focused on what they need to do to pass the exam, not on learning the broader set of writing skills required for success beyond the exam context. As one English 30-1 teacher put it, "Students only see the end result. 'I need to get this mark to get into university,' but they forget that once they get into university they need to stay there" (Slomp, Graves, & Broad, 2014, p. 548).

It is no wonder, then, that the *Alberta Assessment Study* also reported significant gaps between public and professional opinion on the diploma exam program (Webber, Aitken, Lupart, & Scott, 2009).

Ultimately, however, the IUA must weigh evidence regarding the intended and unintended consequence of the English 30-1 diploma exam program on students. In this case, the question that begs answering is: Do the benefits for students of acquiring and using reliable metadata outweigh the negative consequences for students that result from the limitations in the test that was used to collect that data? Examining the chain of inferences constituting the IUA argument for the English 30-1 exam suggests that at best the answer to that question is uncertain. Alberta Education's own documentation raises significant questions about the IUA argument in support of this exam.

5.0 Validity's Challenges in Guiding Ethical Practice

The above analysis illustrates how robust Kane's (2013) IUA model of validation is. Even this simplified application of the model illuminates serious questions about the quality of Alberta's English 30-1 writing exam. Examined in light of the six dimensions of our theory of ethics, the strengths and limitations of the IUA model become clear. A strength of the model is its emphasis on construct representation and disaggregation of data that are called for in support of the generalization inference and the decision inference. Weaknesses of the model are that it does not address the issue of opportunity to learn anywhere within the inferential chain. More seriously, especially within the writing assessment context, commitment to ensuring strong support for the scoring inference undermines a focus on both socio-cultural perspectives and access. Standard practices and protocols for achieving consistency among raters undermine a socio-cultural perspective by implying there is one objective reading of a student text (Slomp, 2015). Similarly, attention to the scoring inference undermines attention to access because of its focus on a one-size-fits-all context in which students must write the exam. Perhaps the most significant issue with the model is its inattention to the issue of justice highlighted by the fact that Alberta's exam program has been in use for 34 years in spite of the gaps in the validity argument needed to justify its use. Attention to justice should compel action to remedy the weaknesses of both the exam and its program of

validation.

The model, itself, however, has a number of additional limitations that reflect issues with validity theory as it is more broadly conceived. Primary among these is that it conceptualizes core values of fairness, validity, and reliability independently of one another. As a consequence, each of the evidential categories discussed above, because they exist as silos, can be debated one-by-one and can be traded off against one another. This renders an enormous burden on all parties resulting in endless debate and limited action.

Alberta Education's failure to execute and publicize a robust, integrated program of validation for its English 30-1 diploma exam program reflects a broader malaise within the measurement community (Brennan, 2006). The most recent *Standards* (AERA, APA, NCME, 2014), for example are woefully inadequate for providing guidance on how to develop a strong integrated validity argument. The four paragraphs dedicated to this issue (pp. 21-22) are notably vague (an issue made all the more significant given the enormously precise nature of the document as a whole):

[A] test interpretation for a given use rests on evidence for a set of propositions making up the validity argument, and *at some point* validation evidence allows for a summary judgment of the intended interpretation this is well supported and defensible. *At some point* [emphasis added] the effort to provide sufficient validity evidence to support a given test interpretation for a specific use does end (at least provisionally, pending the emergence of a strong basis for questioning that judgment). *Legal requirements may necessitate* [emphasis added] that the validation study be updated in light of such factors as changes in the test population or newly developed alternative testing methods. (p. 22)

What is clear from this section is that the *Standards* provide clear guidelines for the design of assessment programs, but when it comes to articulating a mechanism for developing an integrated validity argument to test that design, the authors of the *Standards* seem to invite test designers and users to walk away without addressing the key issue: Once the design process has been finished, somebody else is left to appraise its quality and impact. These limitations in the *Standards* make all the more transparent the need for a theory of ethics for the field, one that explicitly calls for an integrated approach to the design and appraisal of writing assessment programs.

It is also clear from the analysis of the English 30-1 diploma exam program that while Kane's framework provides an excellent approach to investigating the soundness of test score interpretation and use, it reflects a significant limitation with validity theory itself: its conscious narrowness. Kane's framework determines the appropriate use of test scores, and it examines the impact of score use on the intended population of test-takers, but it does not address broader questions about the impact of such tests on student development, nor does it address the impact of such assessments on the broader educational system.

6.0 Looking Forward: Ethics as an Integrated Design and Appraisal Framework (IDAF)

Our focus on ethics has been designed to address these limitations of validity theory and to highlight the need to develop a broader framework for examining the design and use of writing assessments. In 2014, along with two of my graduate students, I published an integrated framework for developing and appraising writing assessment programs (Slomp, Corrigan, & Sugimoto, 2014). The framework was designed to integrate concerns for validity, reliability, and fairness into a single set of questions and considerations for the design and use of writing assessment systems.

In the pages that follow, I present the integrated design process I developed out of this framework. This design process elaborates the concept of assessment as research, first articulated by Brian Huot (2002). Assessment as research was developed to counter the technocentrism that has historically guided the design and validation of writing assessments. Similar to Zumbo and Chan (2014), Huot observed that writing assessment design has been limited by narrow adherence to established technologies of assessment (both theoretical and instrumental). In addition to constraining validation practices, Huot argued many established methodologies for operationalizing reliability, validity, and fairness unduly constrain possibilities for writing assessment design.

Assessment as research, Huot argued, is more open-ended, beginning with the assessor's information needs before envisioning designs that will best fills those needs. Other writing assessment scholars have also adopted this approach (Gallagher & Turley, 2012; Adler-Kassner & O'Neill, 2010). The IDAF I am proposing, like the Evidence-Centered Design (Mislevy, Steinberg, & Almond, 2002) and the Design For Assessment (White, Elliot, & Peckham, 2015) frameworks that precede it, begins from this point. Drawing on both quantitative and qualitative research traditions, assessment as research permits a broader approach to addressing concerns for reliability, fairness, and validity. It also offers a process through which problematic assessments, like the Alberta English 30-1 exam, can be rectified.

Huot's (2002) initial heuristic encompassed nine questions that designers of writing assessments should attend to. These questions focus assessment developers on identifying information needs; defining methods for collection, verification, and reporting; and determining methods for examining impact (p. 181). This heuristic provides a solid framework for highlighting the major concerns

assessment developers need to pay attention to when designing assessments. The IDAF model I am advancing extends Huot's framework by expanding the range of questions that must be considered, and by providing clear direction on how to translate the answers to these questions into procedures for assessment design,

I have expanded these nine questions into a more complete heuristic and I have structured it into six reiterative phases that pair a series of questions with a series of actions that correspond with them (see Table 1). In phase one of this process key issues are identified. These include: the information needs motivating the assessment; the audiences for that information; the inferences, decisions, and actions to be taken on the basis of that information; the populations to be affected by the assessment; and the intended and unintended positive and negative consequences that could result from this assessment process. In phase 2 this information is used to identify elements foundational to the assessment's design. Emphasis is on defining the constructs and content domains at the core of the assessment program. In phase 3 the assessment instruments are developed. This work begins with a survey of the information collection methods available. Each potential method is then critically appraised based on how it might effect construct underrepresentation or construct irrelevant variance. In phase 4 the scoring system is developed. Potential scoring procedures are identified and screened for how they affect construct representation, and for how they impact populations of interest. In phase 5 a plan for analyzing assessment results is developed; particular attention is given to analyzing the strength of the scoring, generalization, and extrapolation inferences while also attending to issues of disparate impact. In phase 6 a plan for analyzing assessment consequences that pays particular attention to intended, unintended, positive and negative outcomes is developed and executed.

Table 1: Integrated Design and Appraisal Framework (IDAF).

Phase 1: Define assessment aims, principles of fairness, and context.	
Actions Required	**Questions to be Answered**
• Identify information needs with respect to student learning or program functioning that need to be filled. • Identify the inferences to be made on the basis of this information. • Identify the decisions or actions to be made on the basis of this information. • Identify both who will be affected by this assessment and what voice they will have in the assessment design, score interpretation, and decision-making processes. • Identify who the audience is, what values they hold pertinent to the assessment design work, and how the assessment information should to be presented these diverse audiences. • Identify the positive intended consequences to be promoted through this assessment process, including how resources will be allocated to the least advantaged. • Anticipate the unintended positive and negative consequences that could result from this assessment process.	1. Who are the stakeholders that will be impacted by this assessment procedure? a. What voice will they have in the design, implementation, and use of this assessment? b. What voice will they have in the decisions being made on the basis of this assessment data? 2. How will the development and use of this assessment promote equity of opportunity? 3. What are the information needs about students, teachers, or administrators that need to be filled? a. What do you want to know? i. About or from students ii. About or from teachers iii. About or from administrators 4. What inferences about student learning or program functioning will be made from the assessment information? 5. How will this assessment information be used? a. What decisions or actions will be taken based on this information? 6. Who are the audiences for this information?

Phase 2: Identify elements foundational in the assessment design.	
Actions Required	**Questions to be Answered**
Keeping aims, information needs, and audiences in mind: • Identify and define the constructs to be measured. o Review the research literature pertinent to defining each of the constructs being measured. ▪ Create a map of the construct features. • Identify the domain or domains being assessed. o Review the outcomes in the pertinent curriculum or program documents. ▪ Create a map of the content domain needed for success.	1. What construct or constructs are targeted in the assessment? a. How well is each construct understood? b. How stable is each construct across social, cultural, or racial contexts? 2. How will the construct sample be defined? a. Does the construct sample avoid the issue of construct underrepresentation? 3. What is the content domain being measured? a. How will that content domain be sampled? i. Does the content sample ensure confidence? ii. Does the content sample introduce construct underrepresentation or construct irrelevant variance into the assessment plan?

Phase 3: Develop assessment program.	
Actions Required	**Questions to be Answered**

• Identify the collection methods needed to capture the information required.
  o Review information collection methods appropriate to information needs.
• Generate a list of potential information collection methods.
• From among the data collection methods listed in **Step 2**, develop a system that will capture the knowledge, skills, and attitudes described above.
  o Ensure assessment system allows for multiple sources of information.
  o Map task features onto the construct and content maps developed earlier.
  ▪ What is missing?
  ▪ What is being measured that is not on the map?
  ▪ How will the assessment system allow for measures of inter-topic reliability?
	1. What is the range of approaches you can use to collect the information you want?
	  a. From among this range, which approach (or suite of approaches) best captures this information?
	  i. How completely do these approaches capture the intended domains associated with the construct? (C-Underrepresentation)
	  ii. How cleanly do these approaches capture the intended construct(s)? (C-Irrelevant-variance)
	2. In what ways do the selected approaches introduce irrelevant variance into the measurement process?
	3. How will the approach to collecting this information impact student learning and/or program delivery?
	  a. Considering the populations being assessed, how might the assessment lead to disparate impact?
	4. What form will the information take?

Phase 4: Develop scoring system.	
Actions Required	**Questions to be Answered**

• Develop scoring criteria
  a. Map criteria onto the construct and content maps developed in **Step 2**.
  i. What is missing?
  ii. What is contained in the scoring criteria that is not contained on the maps developed in **Step 2**?
• Develop scoring process
  a. Ensure scoring process involves a method for ensuring consistency (consider range of options from research methods for achieving this).
	1. What are your scoring criteria?
	  a. How do these criteria capture the construct(s) and content domain being measured?
	  b. How will component scores be combined into a total score?
	  c. If a total score is needed, how will sub-scores be analyzed to examine their impact?
	  d. Does your total score either over-represent or under-represent some aspects of the construct or content domains being measured?
	2. What is the range of scoring procedures available to you?
	  a. What are the strengths and weakness of each procedure?
	  b. Which of the available procedures will you use to score student performances?
	3. How will scoring procedures achieve consistency (reduce irrelevant variance) while also supporting construct validity?
	4. How will scoring procedures influence
	  a. Assessment outcomes?
	  b. Student populations?
	  c. District, school, and/or classroom experiences?

Phase 5: Develop a plan for analyzing assessment results.	
Actions Required	**Questions to be Answered**

• Develop a process for analyzing assessment results. Process should enable comparison of results against construct and content maps to determine what assessment results enable test users to say about:
  a. Performance on the test tasks themselves.
  b. Performance against the broader construct/content domain.
  c. Performance beyond the assessment tasks and construct/content domains.
• Develop a plan for considering the disaggregated data:
  a. Plan should identify individuals or populations whose results should receive greater scrutiny.
	1. *Scoring inference* -- What claims about performances on the test items themselves does the information enable you to make?
	2. *Generalization inference* -- Reflecting on construct sample, what claims about performances on the broader construct domain does the information enable?
	3. *Extrapolation inference* -- What claims about performance beyond the specific test context, construct, or set of test tasks does the information enable?
	4. For each of the inferences stated above,
	  a. What is the argument that supports the inference?
	  i. What research evidence supports the argument?
	  b. What are the qualifications to the general argument?
	5. Within the population being assessed, have the populations who might be disparately impacted by the assessment been identified?

iii. Identify whether these individuals or populations are performing better or worse than the norm. iv. Analyze data on assessee's processes, comparing processes across performance groups to determine extent to which performance is tied to construct being assessed. v. Conduct follow-up data collection and analysis as needed.	a. How will test developers ensure each population is represented in sufficient quantity to allow for meaningful analysis of their performance? b. How will the performance of each population be examined and compared? i. Can differences in performance between populations be attributed to actual differences in ability in relation to the construct being measured? c. In cases where differences in performance across populations have been identified, does evidence based on test content, response processes, and internal structure indicate the assessment is measuring the same construct across populations? i. Do differences in the construct being measured undermine the scoring inference, generalization inference, and extrapolation inferences articulated for the assessment? ii. Do differences in the construct being measured undermine decisions made on the basis of assessment results?

Phase 6: Develop a plan for analyzing assessment consequences	
Actions Required	**Questions to be Answered**
• Review intended consequences (positive and negative) for the assessment program. **a.** Develop a plan for analyzing and collecting information on intended outcomes. **i.** Consider how this information collection could be built into the assessment process itself (see **Step 2**). **b.** Develop a plan for uncovering and responding to unintended consequences (positive and negative). **c.** For individuals, populations, and/or the system.	1. Taken collectively, does the evidence gathered indicate that the assessment has achieved the purpose or goals for which it was designed? a. What are the intended (positive and negative) consequences i. For each population affected by the assessment? ii. For the district, school, or classroom serving those populations? 2. Taken collectively, does the evidence provide an understanding of unintended impact, whether positive, negative, or unknown? a. What are the unintended consequences i. For each population affected by the assessment? ii. For the district, school, or classroom serving those populations?

In each section that follows, I provide a description of an assessment design project I have been engaged in with eight teachers in southern Alberta. We have titled this project the Horizon Project. Our design work described below follows the questions posed in Table 1. The following case study sections reflect, as much as possible[2], a real world application of this IDAF.

6.1 Phase 1: Fairness, Aims, and Context

The first phase of this design process focuses on identifying the aims of the assessment program, of situating that assessment program within its ecological context, and of establishing the focus on fairness from the outset of the design work.

6.1.1 Context: Assessing for transfer. The Horizon Project emerged at the intersection of my theoretical work exploring the nature of the writing construct in the 21st century context (Slomp, 2012; Leu, Slomp, Zawalinski, & Corrigan, 2015; Slomp, Graves, & Broad, 2014), and my involvement for the past year in supporting a cohort of new English language arts teachers employed in a southern Alberta school district. Our professional development sessions began by addressing the pressure they were feeling to focus their writing instruction exclusively on preparing students for the provincial achievement tests and diploma exams. To this end, we explored the construct of writing ability—largely focused around Beaufort's (2007) construct model, which described the five domains of metacognitive knowledge expert writers draw on when creating text: Discourse Community Knowledge, Genre Knowledge, Rhetorical Knowledge, Writing Process Knowledge, and Subject Matter Knowledge (see Figure 1). Because our goal was to help students develop expert knowledge about the nature of writing, we used these five knowledge domains as lenses through which we analyzed both Alberta's writing assessment program and each teacher's classroom-based writing assessments. Through this process we determined both sets of assessments were not measuring all of the knowledge domains described in Beaufort's (2007) writing construct. We then discussed potential consequences for student learning and long-term development that might reasonably be anticipated from these flaws in construct representation. We hypothesized that the problems many students face when making the transition from writing in high school to writing in university, college, and the workplace (Beaufort, 2007; Dennihy, 2015; Sommers & Saltz, 2004; Thompson, 2002) could very well be attributed to gaps we were identifying. Research that examines, broadly, writing instruction in American schools, seemed to support this hypothesis (Applebee & Langer, 2011; Kiuhara, Graham, & Hawken, 2009). Applebee and Langer (2011) observed that in spite of a widespread move to process pedagogies, many teachers continue to cultivate dependence in their students by breaking down writing tasks, providing explicit instructions and templates, and doing the deep thinking for them. Our own analysis of writing assignments designed by this cohort of teachers reflected this finding. The types of scaffolding they provided to students limited opportunities to develop metacognitive knowledge about writing, making it difficult for students to learn how to attack new writing tasks independently.

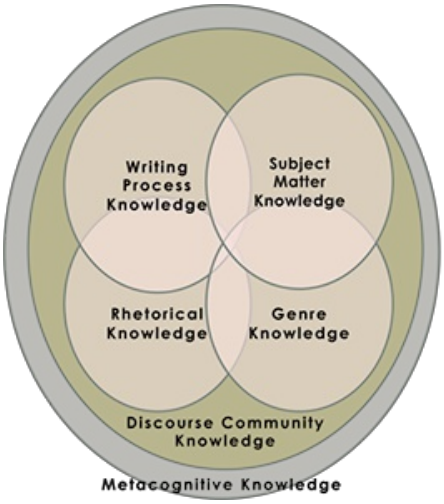

Figure 1: Modified version of Beaufort's (2007) construct model.

Drawing on this hypothesis, we decided to construct, implement, and evaluate a series of writing tasks that would emphasize metacognitive development and that would focus on helping students develop greater independence as writers. Our goal is to foster independence by presenting the writing tasks to students as problems they are required to solve. For this first year of the project, we have designed a pre-intervention task, a concurrent intervention task, and a post-intervention task. The pre- and post- tasks are presented to students without any teacher support or scaffolding. The concurrent task is accompanied by a five-week unit of instruction, during which students are taught a series of metacognitive strategies for interrogating and completing new writing tasks. We have named this a problem-solving model of writing pedagogy (Slomp, Graves, & Broad, 2014).

6.1.2 Fairness. The IDAF process begins by identifying stakeholders who will most be affected by the design and implementation of the assessment system. In this assessment project the primary stakeholders are students and teachers. For students, this project may have long-term implications. If these assessment tasks help foster stronger metacognitive functioning, they have the potential to enhance their development of transferable knowledge about writing, increase their confidence in tackling new writing tasks, improve their resilience in post-secondary education experiences, and enhance their productivity when they enter the workforce. While the immediate stakes—associated with their classroom grades—are low, the long-term stakes may indeed be very high.

For teachers, this project carries much more immediate consequences. Experimenting with new pedagogical approaches is always risky. In jurisdictions where student performance on standardized assessments is closely watched, these risks are enhanced. If this approach does not produce gains in student writing ability, there may be career implications for these teachers. More immediately, involvement in this assessment project enables teachers to develop a much richer and deeper understanding of the writing construct and cultivate greater independence in their students, while also providing them data needed to challenge the restrictive influence of provincial writing assessment programs.

Consequently, teachers and students will both have important voices in the design, implementation, and use of these assessment systems. Teachers have been instrumental to the design of these assessments. They are working on this project as co-researchers. Over the past year the teacher-research team and I collectively explored construct models for writing, limitations in current classroom writing assessment designs, and implications of these insights for our own writing assessment design. During this school year we have met for 11 full day meetings to collective develop and assess the efficacy of this writing-as-problem-solving pedagogy and it attendant assessment program.

Student voice has also been important to shaping these assessment tasks. In September, students in each class completed a small writing-as-problem-solving task: Write a review of a book, movie, video game, or app for teenink.com. Students were given no direction other than the prompt. They had between two and three classes to finish the tasks. In addition to completing the task, students were asked to keep a record of their feeling, questions, and decisions as they navigated this task. After completing the task each class also completed a graffiti wall activity in which they collectively represented their experiences with this task. Additionally, a sample of three students from each class was interviewed about their experience. A section of the interview explored with students what supports they felt they would need if they were given a larger, more complex form of this task.

After all classes had completed this task the eight teachers and I met to identify the challenges students faced. Through this discussion, we refined our major assessment task for the project, and we developed the learning plan that would support student capacity to complete the project.

As the project continues to move forward, teacher and student voices will guide any decisions we make.

The assessment design and implementation process has been created to promote equity by ensuring that included in the project are a range of (a) urban, rural, and distance learning schools; (b) grade levels (6-11); and (c) teachers, from first-year to experienced vice-principals. Additionally, a sampling plan ensuring attention to students from populations of interest has been developed. Analysis of data from this project will examine how each of these factors influence the success of a writing-as-problem-solving pedagogy in fostering student development.

6.1.3 Aims. The information need this assessment system has been designed to address responds to the following questions:

- What transferable knowledge about writing are students (grades 6-12) able to develop and enact?
- Is a problem-solving writing pedagogy able to help students develop and enact transferable knowledge about writing?

The inferences that we hope to draw from this assessment data are as follows:

Inference 1: Students from grade 6-12 are able to develop and enact transferable knowledge about writing.

Inference 2: A writing as problem-solving pedagogy enables the development and enacting of transferable knowledge about writing.

Flowing out of these inferences, the primary decision we will need to make during this project is whether we should continue to pursue developing a writing-as-problem-solving pedagogy. As such, the audiences for this work will be classroom teachers, their students, and eventually, school district and provincial level policy leaders, and the broader writing studies community. In the sections that follow I discuss phases 2-6 of the design process that enable us to provide warrants for these inferences.

6.2 Phase 2: Identify Elements Foundational to the Assessment Design

Phase 2 begins the process of laying a sound foundation for the design of the assessment system. Sound design requires an understanding of three elements: the construct to be measured, the populations to be assessed, and the range of data collection possibilities that will enable collection of required information.

6.2.1 Understanding the construct. Over the years, our understanding of the writing construct has changed as new data and new research traditions inform the field (Camp, 2012). It is critically important that writing assessment programs reflect these evolving understandings. Assessments based on incomplete or out of date construct models can harm students, teachers, and educational systems.

The first step in this process is to identify and define the constructs the proposed assessment will be measuring. Once these constructs have been identified, assessment developers need to conduct an extensive review of the literature on these constructs to ascertain the matrix of knowledge, skills, and dispositions that collectively constitute it. Based on this review, assessment developers need to create a construct map that clearly explicates each element of the construct while also delineating the construct being tested from other related constructs.

The assessments the teacher research team and I are developing are based on Beaufort's (2007) transfer-oriented writing construct. In the predevelopment phase of our project, we worked through a model writing-as-problem-solving task to develop a clear picture of the subset of knowledge, skills, and dispositions associated with each facet of this construct model. The teachers completed this task in groups of three. One group member recorded every question, action and decision the group took as they worked through this task (a form of expert response process modeling). The entire group then inputted their data into a spreadsheet. After the sheet was populated, each line was coded first according to the knowledge domain it was associated with, and then by the curriculum outcome it addressed. They coded this data by comparing each decision, question, and action to brief descriptions of each construct facet before making a judgment about what facet(s) it aligned with. When coding for alignment to curriculum outcomes, they similarly compared each decision, question, and action to the curriculum outcomes to determine which ones they best aligned with. This process provided concrete detail regarding the range of knowledge, skills, and dispositions captured within each knowledge domain described by Beaufort (2007).

Once this map is developed, it needs to be critically reviewed. First, we need to ascertain how well the construct is actually understood. If the body of research into the construct is not broadly based and extensive, assessment developers must question how confident they are in their understanding of the construct. Data generated by tests that measure poorly understood constructs need to be handled with higher degrees of tentativeness. Second, even thoroughly researched constructs, based on a solid, multidimensional body of research, need to be critically examined with a view to how stable this construct is across social, cultural, or ethnic lines. Essentially, construct models need to be examined with respect to each of the populations who will be taking the

assessment. Gould's (1996) critique of IQ testing demonstrates the damage that can occur when thoroughly researched constructs are not critically examined for social, cultural, or racial bias.

In our Horizon Project we are also conducting a critical examination of Beaufort's construct model. Julie Corrigan (Corrigan, 2014; Corrigan & Slomp, 2014) has been leading this aspect of our work. She began this work by abductively coding influential articles articulating key literacy theories (Leu et al., 2013; Street, 2003; New London Group, 1996; Flower & Hayes, 1981; Breuch, 2002). This first round of coding has been used to refine Beaufort's conceptual model, which will provide the foundation for a second, more extensive round of abductive coding of 226 documents (articles, books, book chapters) that are focused on articulating a construct or theoretical framework for writing (Onwuegbuzie & Leech, 2007). Once this work has been completed, each writing-as-problem-solving task will be critically re-evaluated in light of this revised construct model.

This process of critical construct analysis is designed to help answer questions regarding how well the construct is understood. Coding the research with respect to populations involved will also enable us to develop an understanding of how well the construct is understood across racial, cultural and social contexts. It will also enable us to determine how to structure the construct sample for future iterations of our assessment design work.

6.2.2 Curriculum alignment. It is common practice for assessment programs in schools to blueprint their tests against prescribed curricula. These blueprints, however, often fail to first map the curriculum outcomes against the construct features. As a consequence, it is often possible for an assessment program to have strong evidence of content validity, while at the same suffer from issues of construct underrepresentation or construct irrelevant variance. For this reason, the 2014 *Standards* emphasized the dependence of content validity evidence on construct validity. Mapping curriculum outcomes onto the construct map in the manner addressed above enables test developers to understand how and where the curriculum underrepresents the construct and it helps to identify outcomes that introduce irrelevant variance into the assessment context.

In our Horizon Project, the teacher-research team coded their own response processes to the sample writing-as-problem-solving task they completed. They first coded onto construct domains and then onto Alberta's English language arts (ELA) curriculum outcomes. This mapping exercise revealed that every one of Alberta's ELA curriculum outcomes maps onto every domain of Beaufort's construct model (a marked contrast to the more limited curriculum coverage of the English 30-1 exam discussed in §3.3.4).

This process was an important step for the teacher-research team as it gave them faith that any assessment we design using a writing-as-problem-solving model should be able to hit each facet of the curriculum. Consequently they could present our assessment project to their school and district leadership as an integral part of their program, rather than an add-on to their already heavy curricular demands. This mapping process also makes visible how decisions to sample the outcomes might contribute to problems of construct underrepresentation.

6.3 Phase 3: Identifying and Selecting Information Collection Methods

Having developed a strong construct map, one that delineates the knowledge, skills, and dispositions that constitute the construct of interest, assessment development then moves on to consider the range of data collection methods that can be used to capture the required information. Waiting to make design decisions until after the construct mapping has been completed is essential. Because the goal of this process is to develop as clean a measurement of the construct of interest as possible, it is important at this stage to create a list of data collection instruments, to evaluate the strengths and weaknesses of each of those instruments, and to understand what elements of the construct each instrument will be able to capture, as well as what construct irrelevant skills and understandings each instrument introduces into the information collected.

With respect to the writing as problem-solving task, because each of the five knowledge domains are inherently metacognitive in nature, a key focus of this assessment system must be the collection of metacognitive data. This can be collected through a variety of methods including: think aloud protocols, retrospective document analysis, process journals, and interviews. The weakness in all of these methods is that they potentially may distort the metacognitive processes students are in fact engaging in.

At this stage of design it is important to recognize that no assessment is perfect. No design choice is neutral. Each choice will either constrain the ability to capture the construct in some way or it will introduce issues of construct irrelevant variance. Being cognizant about these limitations in advance of making design choices is important because it enables test designers to be more purposeful about their design choices while at the same time being more humble about uses and decisions based on assessment data.

In the Horizon Project we developed three sets of writing tasks. The first writing task was designed to give us a picture of students' independent problem-solving skills prior to instruction. The writing task stated:

> Write a review of a book, film, video game, or app you viewed/read/used this fall to be published on TeenInk.com.

Students were given two to three days to complete the task. They were told they were able to access whatever resources they felt were necessary to help them complete this task. Immediately after the students were given the task, they were asked to first pause and record their initial thoughts and reactions to what was being asked of them. As they were working, students were also required to keep a running record of questions asked, decisions made, and actions taken. After completing the task they were asked to reflect upon what challenges they faced when completing the task, and how they worked through those challenges. While the prompt was different – Write a letter to a Canadian soldier who is on active duty overseas – the process for the final task was the same as for the first.

The major task we designed required students to identify a need in their school or broader community, develop a project or initiative that would help address that need, identify a grant agency/program that would potentially fund that initiative, and then assemble and submit a proposal to that grant agency. In addition to writing this grant proposal, students completed a series of additional tasks: They completed a writing portfolio that documented all the work they did for this project; they conducted an analysis of the alignment between their proposed project and the values and goals of the granting agency they were planning to apply for funding to; they presented their analysis to community members; they interviewed professionals responsible for either writing or adjudicating grants; they completed a visual analogy in which they communicated what they learned about writing from this project, and they prepared either a written or oral explanation of this analogy.

The instructional process that accompanied this assignment guided students through strategies for engaging each of the five knowledge domains—discourse community knowledge, rhetorical knowledge, genre knowledge, writing process knowledge, and subject-matter knowledge—described by Beaufort (2007). Addressing these knowledge domains, students completed the following:

- Researched grant agencies to identify what values and goals the agencies held. This list of values and goals was subsequently used to filter each decision about how to write the grant.
- Examined the purpose of grant proposals.
- Analyzed completed grant proposals, proposal forms, and evaluation criteria identifying typical content, structures, and linguistic features. They then reflected on how these features responded to the values and purposes of the grant agencies.
- Analyzed exemplar documents by completing a *Says/Does analysis* (Elbow, 1998). Once they completed their analysis, students wrote a brief reflection on what rhetorical moves would be useful in their proposal.
- Developed, enacted, and reflected upon a research plan to ensure their proposal included required information and content.
- Developed, implemented, and reflected upon an action list the group needed to complete to finish the proposal.

Once we selected our assessment tools and developed our learning plan, our next task was to map each element of the assessment against the construct map. This process reveals the integrity of the assessment system. It points both to redundancies across tasks and to areas of construct underrepresentation. It also lays the foundation for a plan to integrate data from across assessment tasks. It is important to notice that the tasks as they have been described above are largely open with respect to modality of response. The issue of modality is important. Where possible, multiple modalities should be incorporated into the assessment system; this helps to reduce issues of construct irrelevant variance (to what extent is student performance a result of the modality of presentation). In some cases—especially because we are assessing writing— requiring a written response is necessary; in other cases—for example, when collecting information regarding student metacognitive awareness—prescribing a written modality may distort the information being collected for some students. In the case of this assessment system, the information we will collect will come in the form of oral reports, discussions, visual representations, written work, and multimodal presentations.

Envisioned outcomes for this process are that it enables an analysis of how the development and implementation of the assessment system will influence program development and student learning. The process used to develop our writing-as-problem-solving assessments is designed to influence program development by equipping the teachers involved with a design process that will enable them to develop other high quality assessments independent of the research team well into the future. It also strives to affect program development by building a collaborative team of committed teachers from across an entire school district who will have a deep understanding of the writing construct and the assessment design process. This may facilitate district-wide change in writing instructional and assessment practices over the long term. This in turn should drive improvements in student learning across the district.

The task-to-construct mapping process also provides information regarding the potential effect this assessment design will have on student learning in the near term. We've know for a long time that assessment design drives pedagogical choices, especially when assessments misrepresent the constructs they are designed to measure. Mapping construct features onto assessment tasks provides confidence that teaching to the broad construct will be promoted within the context of this assessment program.

6.4 Phase 4: Developing the Scoring System

Development of the scoring system will take place in two phases. In the first phase, the research team engaged in a process called dynamic criteria mapping (Broad, 2003). Dynamic criteria mapping involves a systematic and collaborative review of source documents, focusing on uncovering what the group values about writing. Once students completed tasks, the research group met

together to read many samples from across the different classes. We then discussed what qualities we valued in each sample. This list of values became the focus of our assessment work. For each task the sets of values differed. Students also contributed to the development of scoring criteria through their analysis of genre features and rhetorical features of sample texts. Developing criteria for evaluating the texts they are creating is an important element of the writing-as-problem-solving process. Student-generated criteria can be used to shape student writer's choices and can be used to create an agreed upon set of criteria for assessing student work.

Phase two of the process of developing the scoring system will involve collaborative discussions with the teacher-research team. Criteria generated by students will be reviewed and compared against criteria generated previously by the teacher-research team and our construct map to determine final criteria.

This process will affect student populations by promoting learning through the process of criteria development. It will influence district and classroom experiences by promoting attention to the full set of construct features identified during the design phase of this work. An initial presentation of our work has been done already at District PD sessions and regional teachers' conferences.

During this iteration of the project, teachers graded samples of student writing independently from one another, largely due to the different timelines each teacher followed on the project. Going forward, however, the scoring process will involve collaborative grading by members of the teacher-research team. Collaborative grading will follow a hermeneutic process (Lynne, 2004; Moss, 1994). Members of the marking team will collectively grade a sample of student work. Differences will be discussed until consensus is reached. After a number of pieces have been collaboratively evaluated, the team will split into smaller groups who will mark a subset of student work. Each marking team will grade a percentage of student work independently. Results of these grading processes will be examined for differences. Differences will be discussed and addressed by the full group.

6.5 Phase 5: Developing a Plan for Analyzing Assessment Results

The next stage in development is to construct a plan for analyzing assessment results. Because we have not fully implemented the project at the time of writing, we have not fully completed this phase. This phase examines the scoring, generalization, extrapolation, and decision inferences in sequence.

Scoring Inference: The scoring procedures for this assessment are appropriate, are implemented correctly, consistently, and free from overt bias.

Scoring procedures for this assessment followed an iterative, hermeneutic process. They were designed to test and support this inference. Our procedures were as follows: The process began with a meeting prior to the school year, where the research team reviewed the construct model for writing at the heart of the writing as problem-solving process. We also reviewed lists of construct features developed by the teachers themselves in the previous year when they were asked to complete similar tasks. During the school year the research team meets to score student work. Following a process of dynamic criteria mapping (Broad, 2003) we developed a list of qualities we valued in student work. After each student writing task has been completed, we will break into two groups for the purposes of designing and testing grading criteria for students' work. Each group will collaboratively review each submission assigned to them and will determine a score. Both groups will mark ten percent of submissions. Scores for these submissions will be compared for inter-group consistency.

At this time, critical questions regarding the scoring procedures will need to be explored. In particular, the values revealed through the dynamic criteria mapping process will need to examined for how well they fit Beaufort's construct model. They will also need to be examined for issues of cultural or linguistic bias. The consensus process, too, will need to be evaluated to determine the extent to which this process privileges perspectives and biases of members of the marking group.

Generalization Inference: The problem-solving writing assessment has been designed so that tasks, contexts, occasions, and raters collectively provide for a consistent measurement of test-taker ability with respect to the transfer-oriented writing construct.

Warrants to support this inference include the following: First, the assessment captures clearly and cleanly a transfer-oriented construct for writing ability. Second, scores received on this assessment tool consistently reflect student ability.

The assessment system has been designed to measure the writing construct as defined by Beaufort (see Figure 1). Students have been asked to complete three writing-as-problem-solving tasks during the semester. Student performance on each task will be compared. Areas of growth will be identified. A random sample of students has been selected to participate in interviews about each of these writing tasks. The purpose of these interviews will be to discuss with students their response processes and any bio-ecological factors shaping their work (Bronfrenbrenner & Morris, 2006; Slomp, 2012; Driscol & Wells, 2012). Once all the tasks have been completed it will be important for the research team to critically examine the response processes for different populations of students. For example, the decision to require students to write Morale Mail letters as the final task needs to be examined in light of

the Mennonite population in this study. Does their religious adherence to pacifism shape their experience of this prompt differently from those of other students? If so, what are the implications for the assessment information generated by these students? How does this impact the strength of the generalization inference?

Each task has been designed to elicit responses related to discourse community knowledge, rhetorical knowledge, process knowledge, genre knowledge, and subject matter knowledge. Analysis of student response processes will be needed to test the assumption that this is indeed the case. Scoring criteria will be mapped to these construct elements.

Extrapolation inference: Writing task scores accurately reflect students' writing ability with respect to the full range of writing tasks in the target domain, including nontest performances in nontest contexts.

Each of the three tasks students will be asked to complete involve a genre students have not written before, a unique authentic audience, and a real-life rhetorical context. While three tasks will not sample the full domain of tasks, the goal of our design is to examine student ability to both analyze new writing tasks and to enact a plan based on this analysis. Student portfolios and assignments, group graffiti walls, and student interviews will all examine whether students are developing greater independence as writers. We anticipate variation in the efficacy of this approach across grade levels, populations, and students. Applying the bio-ecological framework, these variations will need to be examined to determine the source of this variance. Strategies to address sources of variance will need to be developed.

Decision inference: Adopting a problem-solving pedagogy and assessment model for elementary and secondary school writing classrooms will enable students to develop transferable understandings about writing.

Evidence to test this inference will be collected through a series of writing-as-problem-solving tasks: pre-, post-, and concurrent intervention. Response processes across the three writing tasks will then be compared to determine the degree to which students are developing transferable understandings about writing.

Keeping the principle of fairness in focus, it is important to test each of these inferences against populations of interest. With respect to the scoring inference, it is important to check for any signs of overt bias toward populations of students. For example, in our initial design discussion, the research team has been attending to questions of how do we ensure that students with different cultural/linguistic backgrounds (German-speaking Mennonite, ESL, First Nations) are able to best represent their metacognitive knowledge. Multi-modal response options were developed. At the scoring stage we will need to conduct an analysis of scores disaggregated by population to see if this reveals significant gaps between populations. If differences in performances are found on some facets of the overall tasks, a review of the facet for cultural differences/influences needs to be undertaken. With respect to the extrapolation inference, it is important to assess the degree to which selected tasks marginalize populations of interest on social, cultural or racial grounds. Are there writing tasks beyond those sampled that reflect a broader, more diverse set of social, cultural, or racial experiences? This series of analyses is dependent on the strength of the sampling plan developed in phase one of this process, and on the results of the analysis of the disaggregated data. If differences between populations are found an analysis of the source of those differences could explore evidence based on test content, response processes (students and markers) and internal structure, and ecological context.

6.6 Phase 6: Develop a Plan for Analyzing Consequences.

The final stage in the development and review process involves examining positive and negative intended and unintended outcomes of the assessment. While it is true that at each stage of this process, concern for the consequences of design and use have been paid attention to, it is important to also focus in this issue as a specific topic of study, one that pulls together the information collected throughout the design and implementation process. With respect to this Horizon Project, a number of data collection methods will be used, including: teacher learning logs, classroom observations, document analysis of individual and collective student work, and student interviews. Collectively these methods will be used to determine if intended positive consequences are being realized and to explore what unintended negative and positive consequences might be derived from this process. If negative consequences are discovered through this process, a plan for addressing or mitigating those consequences will be developed in collaboration with the teacher-researcher team and their students.

7.0 Conclusion: Challenges and Complexities

The IDAF enacts *sociocultural perspectives* by recognizing that construct models are socially constructed and situated. As such it requires assessment developers and users to critically examine the constructs that inform assessment design. It also emphasizes the important of *maximum construct representation* by compelling assessment developers to draw clear construct maps that link assessment tasks to construct elements and to curriculum outcomes.

The IDAF attends to the issues of *access* and *opportunity to learn* by emphasizing the importance of developing robust sampling

plans in advance of assessment design and implementation. It positions the *disaggregation of data* as a key tool for illuminating challenges to equity in both the development and the measurement of student learning. It calls for assessment developers and users to analyze assessment programs with respect to how effectively they enable students to demonstrate their standing on the constructs of interest, how meaningfully they engage students on associated curriculum outcomes, and how well they attend to the bio-ecological contexts that shape human development and that influence student potential to learn.

The IDAF addresses issues of *justice* by insisting on the critical appraisal of all aspects of an assessment's design and use. It demands that issues raised through this process be made transparent and, wherever possible, be addressed.

What must be clear from the presentation of these two case studies is that no assessment program will ever be perfect. The scope of challenges and issues is simply too great. As White, Elliot, and Peckham (2015) observed,

Understanding program assessment as an ecology reminds us that we are involved in complexities we both do and do not understand. Recognition of the limits of knowledge leads us to a fundamental belief: only an informed instructor, watching a student develop over time, can hope to make a valid claim about the totality of the writing ability of that student. Such a fundamental premise of program assessment will lead to the humility required if meaningful inferences are to be drawn from the information we collect. (p. 32)

Acknowledgements

This Special Issue has been a wonderfully rich collaborative experience. A big thank you to Bob Broad, Ellen Cushman, Mya Poe, and Norbert Elliot for the rich dialogue and fine feedback. Thanks also to Diane Kelly-Riley and Carl Whithaus, editors of JWA, for your support of this SI from the very beginning. Thank you also for your excellent editorial work. A most profound note of appreciation to the wonderful, dedicated teachers—Crystal Hegadus, Jaimie Vanham, Kacie Neamtu, Keith Miller, Lindsey Hagen, Rita Leask, Sean Dupuis, and Taylor Burke—who I have had the privilege of collaborating with on our Horizon writing project.

References

Alberta Education. (2004). *Diploma examinations program*. Retrieved from http://www.learning.gov.ab.ca/k12/testing/diploma/dibgib/examinationprogram.asp

Alberta Government. (2015). *Diploma Examination Multiyear Reports*. Retrieved from https://education.alberta.ca/media/1626647/diploma-multiyear-province-report-graph.pdf

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. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Applebee, A. N., & Langer, J. A. (2009). EJ Extra: What is happening in the teaching of writing? *English Journal,98*(5), 18–28.

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

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.) *Construction vs. Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (pp. 1–27). Hillsdale, NJ: Erlbaum.

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.

Breuch, L. M. K. (2002). Post-process "pedagogy": A philosophical exercise. *JAC: A Journal of Rhetoric, Culture, and Politics,*

22(1), 119–50.

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

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology, Vol. 1: Theoretical models of human development* (6th ed., pp. 793–828). New York, NY: Wiley.

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

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.

Corrigan, J. (2014, April). *A Conceptual Model of New Writing: Beaufort's Writing Knowledge Domains 2.0.* Paper session presented at the AERA annual meeting in Philadelphia, PA, USA.

Corrigan, J., & Slomp, D. (2014, May). *Confronting boundaries in writing assessment: Transforming knowledge for a digital age.* Paper presented at the Canadian Society for Studies in Education in St. Catharines, ON, CAN.

Cronbach, L. J. (1970). Mental Tests and the Creation of Opportunity. *Proceedings of the American Philosophical Society, 114*(6), 480–487.

Dennihy, M. (2015). "Forget what you learned in high school!": Bridging the space between high school and college. *Teaching English in the Two Year College*, *43*(2), 156.

Dryer, D. B. (2016). Appraising Translingualism. *College English*, *78*(3), 274–283.

Elbow, P. (1998). *Writing with power: Techniques for mastering the writing process* (2nd ed.). New York, NY: Oxford University Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Gallagher, C., & Turley, D. (2012). *Our better judgment: Teacher leadership for writing assessment.* Urbana, IL: NCTE.

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

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

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

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

Kiuhara, S. A., Graham, S., & Hawken, L. S. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, *101*(1), 136.

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.

Leu, D. J., Kinzer, C. K., Coiro, J. L., Castek, J., & Henry, L. A. (2013). New Literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell, *Theoretical models and processes of reading* (6th ed.) (pp. 1150–1181). Newark, DE: International Literacy Association.

Leu, D. J., Slomp, D., Zawilinski, L., & Corrigan, J. (2014). Writing from a new literacy lens. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing* research (2nd ed.). New York, NY: Guilford.

Lynne, P. (2004). *Coming to terms: A theory of writing assessment.* Logan: Utah State University Press.

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496.

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

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*(1), 60–92.

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.

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.

Slomp, D. (2015). Writing assessment in six lessons–from "American Idol". *Phi Delta Kappan, 96*(5), 62–67.

Slomp, D., Graves, R., & Broad, B. (2014). (Re-)Mapping the system: Toward dialogue-driven transformation in the teaching and assessment of writing. *Alberta Journal of Educational Research, 60*(3), 538–558.

Slomp, D., Corrigan, J., Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments. *Research in the Teaching of English, 48*(3), 276–302.

Slomp, D. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing, 17*, 81–91.

Slomp, D. (2008). Harming not helping: The impact of a Canadian standardized writing assessment on curriculum and pedagogy. *Assessing Writing, 13,* 180–200.

Slomp, D. (2007). *Trapped between paradigms: Composition pedagogy in the context of a twelfth grade standardized writing assessment*. Unpublished doctoral dissertation, University of Alberta, Edmonton, AB, CAN.

Sommers, N., & Saltz, L. (2004). The novice as expert: Writing the freshman year. *College Composition and Communication, 56*(1), 124–149.

Staples, D. (2012, Nov, 19). Educational assessment debate. *Edmonton Journal*. Retrieved from http://edmontonjournal.com/news/local-news/educational-assessment-debate-standardized-tests-are-necessary-for-both-accountability-as-well-as-for-consistency-in-reporting-student-achievement-educational-expert-argues

Street, B. (2003). What's "new" in New Literacy Studies? Critical approaches to literacy in theory and practice. *Current Issues in Comparative Education, 5*(2), 77–91

Thompson, T. (Ed.). (2002). *Teaching writing in high school and college: Conversations and collaborations*. Urbana, IL: NCTE.

Webber, C. F., Aitken, N., Lupart, J., & Scott, S. (2009). *Alberta Assessment Study: Final Report.* Edmonton, AB: Government of Alberta.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Logan, UT: Utah State University Press.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1–10.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*, 37–53.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY:

Springer.

Notes

[1] I have chosen to focus on the Alberta context because my experiences preparing students for the English 30-1 exam first prompted me to question the ethics of writing assessment. Additionally, Alberta's English 30-1 exam program is a high-quality example of the timed, impromptu model of standardized writing assessment used in multiple contexts throughout the world.

[2] Some of this work has already begun, while other elements of will not be completed until the end of the 2015/16 school year.

Author Biography

David Slomp is an Associate Professor of Literacy and Assessment in the Faculty of Education at the University of Lethbridge. His appointment as the University of Lethbridge Board of Governor's Teaching Chair begins in July 2016.