English Teaching: Practice and Critique December, 2005, Volume 4, Number 3
http://education.waikato.ac.nz/research/files/etpc/2005v4n3art8.pdf pp. 141-155

Teaching and assessing language skills: Defining the knowledge that matters

DAVID H. SLOMP
University of Alberta

ABSTRACT: A goal of this double issue of English Teaching: Practice and Critique *is to collectively consider what we mean when we talk about knowledge about language. How have our understandings changed over time? What are the implications of these new understandings for pedagogy in the field of language teaching? These are necessary and important questions. This article, however, does not attempt to address them. Rather, it focuses on the power of standardized assessment in language education and on its implications for the discussions contained within this journal. Central to this paper is the argument that standardized language assessments are resistant to change, rarely integrating new understandings of language into assessment designs. This reticence in turn limits advances in pedagogy. Language theorists and educators are therefore compelled to advocate for assessment reform. Drawing on a study of government-mandated writing assessment and its impact on Grade 12 academic students in Alberta, Canada, this article demonstrates how poorly developed standardized assessments curtail teaching and learning. The article concludes with a discussion of validity theory and its implications for test design, demonstrating how validity research can be used to ensure that standardized language tests value and support new understandings of language theory.*

KEYWORDS: Language testing, consequential validity, construct validity, ethics, pedagogy, writing.

> *Instruction: Select the response that best answers the question. You may only circle one letter.*
>
> **Whose knowledge about language counts most in the Language and Literacy classroom?**
>
> **a. the teacher's**
> **b. the students'**
> **c. the language and literacy researcher's**
> **d. the high-stakes assessment designers'**
> **e. the curriculum developer's**
> **f. the cognitive psychologist's**
> **g. none of the above**
> **h. all of the above**
> **i. some of the above**

I have somewhat facetiously framed the introduction to this article around a poorly written, multiple-choice, test question to draw attention to the key question this article explores, to a range of its possible answers, and to the context within which this question gains its importance.

I would argue that currently in North America, given the increasing prevalence and power of high-stakes testing, the above question would be answered with the letter (d) – the assessment designer. Testing, after all, is the action through which we state

most clearly and publicly what knowledge is either valued or not valued within a system of education. Though one would expect a clear alignment between what is valued in the educational community as a whole and what is valued by a government mandated test, this desired alignment is not always realized. When the language test, the Language Arts curriculum, and the theories of language which underlie them conflict with one another, espousing different views of language, students and teachers are compelled to decide which view they will adopt for themselves. And when a test score is used (even partially) to determine college entrance, graduation eligibility, or even school funding, the outcome of this choice is fairly predictable.

In Canada, curriculum development and language and literacy research are effectively pursued through collaborative networks such as the Canadian Language and Literacy Research Network (www.cclrnet.ca) and the Western and Northern Canadian Protocol for curriculum development (www.wcp.ca). Our assessments, however, are developed in a much less collaborative manner. Too often government-mandated tests are imposed on resistant educators. Over the past fifty years, educators have had some success in pushing for improvements to language testing. However, significant issues remain unresolved (Yancey, 1999; Hamp-Lyons 2002). This progress, though, has been far from uniform. For example, George Hillocks (2002) reports on five writing assessment programs in the United States which collectively demonstrate a significant range in assessment quality across states. In Canada, many of the most recent advances described by Yancey (1999) and Hamp-Lyons (2002) have not been implemented in provincially mandated writing assessments programs. Given this reticence, new knowledge and understanding of language (that is, the material contained in this double issue of *English Teaching Practice and Critique*) often does not find itself represented in current language tests. As a result, such knowledge is implicitly devalued and deemphasized within education systems. As long ago as 1975, James Britton had demonstrated that high-stakes writing tests contributed to a narrowing of pedagogical focus (Britton, 1975). Teachers who were preparing students to write the high-stakes tests understood what knowledge about language the test was valuing and what knowledge it was devaluing; they tailored their instruction accordingly.

Educators, researchers and language theorists need not remain passive victims of the assessment industry. George Hillocks (2003), in his paper "Fighting Back: Assessing the Assessments", argues that educators must confront assessment issues head on. By providing a list of questions educators and researchers can use to interrogate high-stakes writing tests, he in part lays the foundation for such an attack. His article, however, fails to provide a framework through which a coherent body of research related to the assessing of assessments can be developed.

This article will explore this issue in two parts. Firstly, it will illustrate the potential problems that arise when tests and theory do not align. This section will be based upon qualitative research conducted with three, Grade Twelve teachers and their students in an academic English course in Alberta, Canada. The second section of this article will explore important aspects of validity theory and its implications for teachers, researchers and theorists who wish to challenge the status quo.

DISPARATE VALUES IN LANGUAGE EDUCATION

My current research investigates the impact that high-stakes writing assessments exert on the teaching of writing in Alberta, Canada. The study is designed according to a multi-method research design which utilizes three case studies and two surveys. The case study component of this research involved working with three teachers and their Grade 12 academic English classes (English 30-1). I collected data through a series of interviews with teachers and students, through the collection of writing assignments, and through classroom observations.

At the end of the English 30-1 course, students are required to take two government-mandated exams. The first exam is a reading comprehension exam, the second is a writing exam. Each exam is worth 25% of a student's final grade in this course. In Alberta successful completion of this course is a requirement for admission to any of Alberta's three major universities.

Alberta's Academic Grade 12 English writing examination

The basic format of Alberta's English 30-1 writing exam is as follows: Students are given a maximum of three hours to complete two essay questions. These questions are linked thematically. The first question, the *Personal Response to Text Assignment*, is designed to stimulate student thinking for the second question, the *Critical/Analytical Response to Texts Assignment*. The exam permits students to respond to the questions from either a personal, critical or creative perspective. As well, the exam permits students to express their ideas in any form that they deem appropriate to the ideas they wish to express.

Personal Response to Text Assignment. The suggested time for students to complete this assignment is between forty-five and sixty minutes. Before writing, students must read through four pages of print text and visual text provided. These texts are followed by a prompt which places them into context or which focuses the students' attention to elements of the text that are most relevant to the writing prompt that follows. The prompt in the January, 2004 version of the exam read: "What do these texts suggest to you about the significance of our memory of the past? Support your idea(s) with reference to one or more of the texts presented and to your previous knowledge and/or experience (Alberta Education, 2004, p. 7). Below the prompt, was a series of reminders for students:

- Select a *prose form* that is appropriate to the ideas you wish to express and that will enable you to effectively communicate to the reader;
- Discuss ideas and/or impressions that are meaningful to you (p. 7).

Four pages for planning and four pages for writing were provided.

This section of the exam is generally graded according to two, five-point analytic scales. The first scale, *Ideas and Impressions*, is focused on the quality of students' ideas, reflection and exploration of the topic. It also focuses on how effectively they support these ideas, reflections and explorations. *Presentation*, the second scale, focuses on:

- The effectiveness of voice and its appropriateness to the intended audience of the prose form the student has chosen;
- The quality of language and expression;
- The appropriateness of development and unifying effect to the prose form (Alberta Education, 2004, p. 70).

Markers are prompted to consider the proportion of error to the complexity and length of the response. The scale is somewhat relative; within different contexts certain types of errors will be scored more severely than others.

Critical/Analytical Response to Texts Assignment. The suggested time for this assignment is between one and a half and two hours. Students are provided with a writing prompt. The prompt for the January, 2004 exam was as follows:

> Consider how the significance of memory of the past has been reflected and developed in a literary text or texts you have studied. Discuss the idea(s) developed by the author(s) about the significance of our memory of the past (Alberta Education, 2004, p. 8).

Students were also provided with a series of reminders for planning and writing:

- When considering the works you will discuss, choose texts that you know well, that are meaningful to you, and that are relevant to this assignment. Choose from those texts that you have studied in your high school English classes.
- Carefully consider your controlling idea or how you will create a strong unifying effect in your composition.
- You may choose to discuss more than one text.
- As you develop your ideas, support them with appropriate, relevant and meaningful examples from literary texts (Alberta Education, 2004, p. 8).

Students were provided with ten pages for writing and ten pages for planning.

The assignment is generally marked using five, five-point analytic scales: (a) *Thought and Detail* is focused on how effectively the students' ideas relate to the assignment and on the quality of the literary interpretations and understandings the students develop; (b) *Supporting Evidence* is focused on the selection and quality of evidence and on how well the supporting evidence is integrated, synthesized and/or developed to support the student's ideas; (c) *Form and Structure* is focused on how well the student's organizational choices result in a coherent, focused, shaped and concluded discussion and in a unifying effect or a controlling idea that is developed and maintained; (d) *Matters of Choice* is focused on how effectively students' create voice through their use of diction, syntax, and other factors; (e) *Matters of Correctness* focuses on the student=s correct use of sentence construction, usage, grammar and mechanics. Markers are required to consider the proportion of error in relation to length and complexity when assessing Matters of Correctness.

What knowledge about language does the exam value?

To determine what the exam values, one must look at the content, the scoring mechanisms and the structure which collectively constitute the exam. An analysis of the content and scoring mechanism reveals the following: The exam values knowledge about language structure – the structure of ideas, of paragraphs, of sentences. The exam also values knowledge about language as a tool through which one communicates ideas. To this end, it values idea formation and support, and it values the creation of appropriate voice. Knowledge about voice is complex requiring knowledge about diction, syntax and punctuation.

An analysis of the exam's structure also reveals the knowledge and skills valued by the exam. Primary among these values is one's ability to generate, organize and effectively present one's ideas within tightly controlled timeframes. As a consequence of this emphasis on time controls, the exam also seems to place a value on one's ability to work effectively under pressure.

It is also important to think of the exam in terms of what it does not value. Given its short timeframes, the exam neither values knowledge about, nor the skill involved in, developing an effective writing process. It is impossible for students to work through an effective recursive writing process while completing two essays in three hours. The exam values a limited form of writing process; in its reminders to students it merely calls for planning, drafting and polishing. The exam ignores substantive revision as an element of writing process. Its scoring criteria, too, do not measure writing process.

THEORY, TEACHING, AND PROFESSIONAL WRITING

An analysis of writing by leading theorists and expert teachers of writing (Peter Elbow, 1981; George Hillocks, 1987; Donald Murray, 1968, 1990; Kim Stafford, 2003; and William Zinsser, 1988) and professional writers (Margaret Atwood, 2002; Joan Didion, 1994; Annie Dillard, 1989; Stephen King, 2000; George Orwell, 1994; and Carolyn See, 2002) reveals a consensus in terms of what skills and knowledge about writing is valued by leading figures in this community. The skills and knowledge most emphasized by these individuals are related to motivation to write and the development of an effective writing process. These professionals conceptualize the writing process as a recursive two-stage process containing a creative and a critical stage. Each stage, they agree, requires a different method of thinking. The first requires a purely creative orientation; the second a critical one. They agree that the creative stage of writing is marked by confusion and structural chaos. The second stage of the writing process is marked by a search for coherence. It involves repeated rewriting, editing, shaping and polishing. They also value revision – which involves a critical appraisal of thinking and organization – as being an important element of writing process. And, they value writing as a tool through which one can explore new ideas and come to understand one's experiences.

Divergent values in assessment and theory

The exam and the theory certainly align to some degree with one another in terms of what they value. Both the exam and the theory value knowledge about organization, structure, idea development and the use of mechanics. However, it is in the area of process that the exam and the theory differ significantly. The theory calls for a two stage process, one which separates the creative and the critical stages. Peter Elbow clearly defines the rationale for this separation:

> Writing calls on two skills that are so different that they usually conflict with each other: Creating and criticizing….You'll discover that the two mentalities needed for these two [skills] flower most when they get a chance to operate separately (Elbow, 1981, p. 7).

The exam, however, values a more limited process in which the critical and creative stages are melded together. Elbow argues that this process leads to poor writing. The exam also values the ability to generate material under pressure while the community at large values the ability to cope with and reduce external pressure in order to enhance writing quality.

This divergence of values places teachers in an awkward position. Do they teach to the test which determines 25% of a student's final grade for their English 30-1 course, a course which they need if they want to graduate from high school and attend university? Or do they base their teaching on the consensus of writers, theorists and expert teachers of writing?

The teachers I interviewed felt conflicted on this matter. On the one hand they recognized that the exam in part shared their values in regards to what knowledge about language was important. On the other hand, they recognized that the exam contained some significant flaws. Anne characterized the exam in the following way:

> I think the writing component of the diploma itself is probably the least indicative of what a student can do: it is pressure writing, it is writing out of context. I mean for all the things we teach writing to be, it is not, it is the opposite of everything we want it to be….Is it a measure of what a student is capable of? Yes and no. The strong students who do well under pressure, sure it is. But you know, the majority of us go through life in the mid-range and we suffer from a fair degree of test anxiety. I think that that is a real factor for a lot of kids, so in that sense I don't think it is a fair measure of what a kid is able to do.

Ironically, though Anne was highly critical of this exam, her teaching practice was significantly influenced by it. Anne's major writing assignments were modeled on the exam questions, marking guides and, to a large extent, format. She recognized the tension between her critique of the exam and her teaching practice which she explained as follows:

> What is my goal? As a teacher, is my goal for the kids to have fun, and think "English 30 was the best year, we had so much fun, it was great", or to say "I was really well prepared for my exam. My teacher did her darndest to make sure that I wrote that exam and that I did well on the exam"? I think it requires an essential shift in thinking where we go. My responsibility to the students is to make sure that

> they do well on that exam, and also to their parents and to myself and the administration, you know, to the school basically. A large part of my job is to make sure that those kids do well on that exam, and that is why I had to change. I mean, the principal the first year I was here said, "Don't worry about the exam, don't even teach to the exam, just teach the class, make the kids enjoy it." Oh yeah, sure, that really bit me in the butt, the next year. So I really changed. I had to….Valid or not, the darn thing still exists, and students are going to have to write it. I mean, the only way you could eliminate the tension is to eliminate the exam, and that is not going to happen, that is not going to happen.

While Anne and Brian based their teaching of writing very closely on the type of writing and process expected in the exam, the third teacher I worked with, Heather, did not base her teaching on the demands of the exam. However, toward the end of the school year she dedicated approximately 500 minutes of instruction (some in class, some in seminars scheduled outside of class) directly on preparing students for the exam. In assessment circles, this approach of directly teaching to the exam is seen as unethical (Lane, Park & Stone, 1998). However, I would argue (Slomp, 2005) that such practice is a consequence of poor exam design. If exams were designed to reflect the full range of what is valued in the field, such teaching to the flaws in the test would not be necessary. Heather, reflecting sentiments similar to Anne's explained it this way:

> The parents and the students need to feel assured that if you want to give them this beautiful [methodology] they are still going to have the strategies they need for the exam, they are going to be well aware of what is going to be in there: how to do it; what to take into it; how to prepare. They've got to know that, otherwise you are saying I don't have any responsibility toward that at all. But you do, it is a professional obligation, it is a community obligation. Those parents expect that to happen.

Clearly, my research participants struggled with the tension between the demands of the exam and the demands of pedagogical theory in relation to teaching writing. Ultimately, they recognized that this tension could not be overcome and were forced to make a choice. Given the high-stakes associated with this exam, these educators saw preparing students for this exam, regardless of its flaws, as being an important professional responsibility. These teachers understood implicitly that the knowledge about language which matters in Alberta's K-12 education system was not necessarily the knowledge they had learned to value in their teacher training courses, or in their experiences with writing and with writing theory. Rather, they understood that the knowledge which truly mattered was the knowledge that was assessed on the government-mandated exam.

While understanding the tensions that divergent values place on teachers, it is even more important to recognize the impact of such tensions on student learning. To this end, I interviewed ten students from the three classes I had observed. During these interviews we discussed their experiences with the exam, their experiences preparing for the exam, and their experiences as developing writers. These students recognized the importance of effective writing process in terms of its ability to enhance the quality of their writing. However, when asked whether or not they engaged in a recursive writing process, eight of the ten students I interviewed said, "no". They described a writing process that was limited to the type of process that the exam

demanded of them. They planned, drafted, and then engaged in word or sentence level editing. Major revision or major reconceptualization of ideas did not enter into their writing process. One of Heather's students described his process on a major assignment (they had to write an autobiographical poem) for the course as follows:

> **James**: One thing that really helped [get me started on the assignment] was the first thing that was kind of mandatory. It helped give us a kick in the pants. We had to do research on stuff from our childhood. And that got ideas flowing and just the nostalgia got my mind thinking. That was fun….As I was doing that I was writing down the ideas so I could use them later. After I had all those listed down I just started – starting was the hardest part—once I thought of a line (like we had to have an anchor line which is the main idea), I started with that and then I started throwing in the ideas that I had listed, and as I expressed them I filled up the page and then the ideas kind of flowed and the entire thing just kind of came out once I was looking at the idea I had written down.
>
> **David**: OK, did the ideas come while you were writing or before you did the writing?
>
> **James**: While I did the writing. When I first start I have no idea what I am talking about.
>
> **David**: So you develop the idea while you do the writing. Once you've got that first draft, or that first go-through done, do you go back through it at all or is it pretty much finished?
>
> **James**: I pretty much go through it to make sure I didn't do any spelling or grammar errors. But usually when I am writing I don't like to change my ideas because I am in a completely different mind set than when I was writing it, because my mind is completely different about five minutes after I completed writing it. So I am thinking I just will go through it, I don't want to edit it too much because then it usually ends up sounding like my ideas weren't flowing as well , so I will just make sure it is grammatically correct.
>
> **David**: Would you say that that process is similar or different from the process you use for essays?
>
> **James**: I use the same process.

Clearly James does not follow an involved process when completing his writing: he spends time generating ideas, drafting and then engaging in surface level edits. James' description of his writing process is very similar to the other students' descriptions of their writing processes. While the students I interviewed were learning to value writing process in a theoretical sense they were actually engaging in a limited from of process, a form supported by the exam but not by professional writers or expert teachers of writing.

It may be unfair to blame students' poor writing process on the exam itself. But an analysis of similar government-mandated exams at Grade 3, 6, and 9 levels reveals that Alberta Education – the government branch responsible for elementary and secondary education – consistently puts forward the same message regarding what it values and what it does not value in terms of students' knowledge about writing. It seems that in practice, the students I interviewed and observed have adopted this value system.

Over the past forty-five years our understanding of writing process and its importance for developing effective writers has grown (Murray 1968, Emig, 1971, Elbow 1981, Calkins, 1994). Within the field, consensus over methods for teaching process has

largely been developed (Coles & Volpat, 1985). However, in spite of this consensus among theorists and educators, writing assessments have been largely resistant to change. The experiences of the teachers who participated in my research suggest that high-stakes exams with limited validity in turn limit the instruction that occurs in classrooms. This reality for the field of teaching writing holds important lessons for those engaged in this journal's current discussion of "what counts as knowledge about language?" If we want the ideas contained in this journal to influence pedagogy, we must engage with the government-mandated assessments which publicly declare whether or not the knowledge and skills we values as theorists and teachers will in fact be valued by our systems of education as a whole. The second section of this article explores one method through which this engagement can be accomplished.

VALIDITY THEORY: HOLDING TESTS ACCOUNTABLE

The current mania for testing in North American schools can be attributed to an increasing desire for accountability in education. Current systems of accountability are largely one-way, or top down. Those in government set goals, priorities and funding levels and then devise mechanisms to ensure that the goals and priorities are being met. The Alberta Teachers' Association is currently advocating for a two-way model of accountability, one in which governments and other high-level stake-holders can also be held accountable for their actions within the educational system (Alberta Teachers' Association, 2005). Validity theory can be used as an important element of this two-way model of accountability. It is a powerful tool that teachers, researchers and educational theorists can use to ensure that the values of the educational community at large are in fact reflected in the assessment practices which characterise our systems of education.

Construct validity

Owing to the implementation of a new, high-stakes literacy test in Ontario (Canada's most populous province) there has been a recent growth in discussions related to literacy assessment issues in Canada (Kearns, 2005; Murphy, 2005; Smith, 2005; and Slomp, 2005). While this discussion raises important questions about literacy testing in Canada, the majority of this research is not tied to validity theory and can therefore more easily be dismissed by those involved with the design of assessment systems. Validity theory, however, carries within it ethical and legal obligations that assessment specialists may not ignore. For this reason, research into testing must be tied to test validity if it seeks to make an impact on what is tested, how it is tested and, by implication, what knowledge and skills are valued.

The core element of test validity – the degree to which a test measures what it purports to measure – is the construct: the theoretical representation of the skill or knowledge that the test is attempting to measure. As such, a test's validity rises or falls in accordance with the degree to which a test's scores are a reflection of students' ability in relation to the construct. When we think of constructs it is important to recognize that there are different layers of constructs. On a super-ordinate level would be the construct as it is understood within the scholarly/educational community as a whole. In the case of language education, this would be defined through an answer to the following question: "Collectively as

language researchers and/or teachers what do we agree are fundamental pieces of knowledge and skill that mark effective users of language?" Below this level would be the language curriculum. Based on these theoretical understandings, the language curriculum defines key skills and knowledge students are expected to master during the course of their program of studies. An understanding of the construct as it appears in the curriculum can be developed through a systematic analysis of the curriculum structure, theoretical framework and its learner outcomes or objectives. The final level of construct is the construct as defined by the test. One can understand this construct through an analysis of the test's structure, its content and its scoring criteria. The challenge for test designers is to match as closely as possible the test's construct with the construct as it is captured in the curriculum and as it exists within the literature on language education. When a test fails to achieve this alignment its degree of construct validity becomes questionable.

This failure of alignment generally occurs in two ways: construct under-representation, and construct irrelevant-variance. Construct under-representation occurs when key elements of the construct are not captured within the test construct. In the case of the Alberta English 30-1 writing exam, this is seen in the test's failure to both measure writing process and in its expectation that students employ a truncated writing process. Construct irrelevant-variance occurs when factors extraneous to the construct affect test scores. In the case of the Alberta writing exam, this is seen in the test's insistence on a student's ability to generate error-free (or limited errors) in first draft writing, its insistence on measuring students' ability to write under pressure, and in its measuring of students' ability to generate ideas quickly, under tight time constraints. Students who might be good writers but who struggle to perform effectively under pressure, or who need to work through a multi-staged recursive process to develop effective pieces of writing are likely to be unfairly discriminated against by this exam. Crawford, Helwig, and Tindal (2004) demonstrated that students with learning disabilities performed significantly better in relation to their peers on a high-stakes writing test when the class was asked to complete the test over three days rather than within thirty minutes. This study demonstrates that in the context of time constrained testing, student scores on the exam are not entirely a reflection of students' abilities in relation to the "pure" construct.

Discussions such as the one contained in this double issue are an important first step in validity-based research. As a community of scholars comes together to define its core understandings and to debate the place that new understandings fit within the discipline, we begin to build a collaborative understanding of the constructs that will provide foundations for future language curricula and language tests. To contest existing tests, however, two further steps are needed. First, existing language tests must be analyzed to determine how the tests define and operationalize the constructs they are attempting to measure. Second, this test construct must then be compared to the construct as it is agreed upon in the literature and within the field. Discrepancies between these two constructs in themselves provide an impetus for assessment specialists to redesign their language assessments.

To add strength to any construct-based call for test redesign, one can also examine the consequences such test designs have for students and teachers. My research, for example, demonstrates that as a consequence of Alberta's government-mandated

writing exam program, some students (I hesitate to generalize on the basis of ten students) in Alberta are learning to employ a truncated writing process, one that clearly aligns with the type of process called for by the exam rather than the type of process described by the general construct. My future research will look at a much larger sample of students across a number of provinces, and if this pattern holds, it will clearly provide the basis for an argument that the test is in fact leading students to develop poorer writing skills than otherwise might be the case. Such negative consequences foreground the ethical obligations contained within validity theory.

Ethical implications of validity theory

Messick (1989), a foremost thinker in the field of validity theory, sees test validity as an ethical issue. Simply put, a valid test is also an ethical test. His argument, however, is somewhat more complex.

Messick claims that at the heart of all validity studies is the question, "To what degree, if at all, on the basis of evidence and rationales, should the test scores be interpreted and used in the manner proposed?" (Messick, 1989, p. 5). In formulating this question, he suggests that both the proposed test use and the interpretations of test scores are justifiable on the basis of the construct which under-girds the test. A test of writing ability, for example, should reflect theoretical understandings of the skills needed to write effectively, the process involved in writing effectively, and the criteria which characterize the product "effective writing".

Additionally, inferences drawn from test scores should be justifiable on the basis of the construct the test is designed to tap into. If a student who scores 60% on a writing test is classified by test-designers to be a poor writer, that inference must be attributable to the construct and not to other variables. The theory of writing upon which the test is built should reflect broader understandings of writing theory, so that the inferences drawn from scores derived from the test cannot be called into question. Messick (1989) writes: "Using test scores that 'work' in practice without some understanding of what they mean is like using a drug that works without knowing its properties and reactions" (p. 8). According to Messick, test scores have meaning only in-so-far as they are grounded in the construct. On its own, a 60% score is meaningless. However, the score and the inferences derived from it become meaningful when they can be shown to reflect student ability in relation to the construct. For example, we could define the construct "effective writing" to include the following facets: (a) mastery of multiple strategies for overcoming challenges at each stage of the writing process, (b) effective organization, (c) original and well developed ideas, and (d) highly polished, error-free text. On the basis of this construct, we can provide information to students regarding scores they have received on the test that was designed to measure the construct. A student who uses multiple strategies to negotiate the writing process, who produces polished text, but who fails to develop original ideas and effective organization of material could receive (depending on our scoring system) a 60%. Using the scoring system, an educator should be able demonstrate to the student what the score means in relation to the construct.

Moss (1995) summarizes the implications of Messick's position: "Essentially it [Messick's position] would require that validity researchers provide an *explicit*

conceptual or *theoretical framework* to ground the intended inference and supporting evidence" (p. 6). By implication, this responsibility falls on the test designer as well (Moss, 1992). *The Standards for educational and psychological testing* (AERA, APA & NCME, 1999) expect,

> [t]he construct of interest for a particular test should be embedded in a conceptual framework, no matter how imperfect that framework may be. The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables (pp. 9-10).

In expressing this expectation, *The Standards* certainly reinforce the centrality of the construct and suggest its importance in test development. The phrase "no matter how imperfect that framework may be", however, is problematic. I certainly recognize (and *The Standards* imply) that many constructs in education are complex and that our understandings of them are continuously evolving. I also recognize, however, that test-developers (especially those developing high-stakes tests) have a responsibility to vigorously investigate and comprehensively develop these conceptual frameworks. The phrasing of *The Standards* seems to minimize this responsibility.

Moss (1995) suggests that this decision on wording was probably political. Test-developers suggested that this responsibility would be too great for them to bear alone. She observes that Wiley (1991) attempted to address this concern by differentiating between test validation and construct validation, the latter being more comprehensive than the former. She rightly concludes, however,

> These concerns do not obviate the need for a program of validation research grounded in an explicit conceptual framework and articulated in an integrative argument that justifies (and refutes challenges to) the proposed meaning of test score (Moss, 1995, p. 7).

Her position supports Messick's, and it imposes an ethical burden on test developers: the necessity of developing comprehensive theoretical frameworks in which to embed their tests. It is exactly this emphasis on construct-based, score meaning that ties ethics to validity. Messick (1989) writes:

> One implication of the…formulation is that both meanings and values, as well as both test interpretation and test use, are intertwined in the validation process. Thus, validity and values are one imperative, not two, and test validation implicates both the science and the ethics of assessment (p. 26).

The ethical obligations are clear: Assessments must be valid, based upon a sound understanding of the construct being tested; invalid assessments must be redesigned to better measure the construct; in the process of (re)design, assessment designers must ensure that their test constructs reflect these same constructs as agreed upon by experts in the field. A failure to abide by these obligations places a test designer in a breech of ethics.

Assessment specialists cannot be expected to understand language theory in as full a manner as language theorists or language teachers. Without effective support,

assessment designers will continue to struggle in their attempts to develop language assessments. By implication, then, validity theory places an ethical obligation on teachers and researchers: we must engage with assessment specialists and we must assist them in understanding what knowledge and skills we value, and for what reasons. We must also assist them in developing methods through which these skills and knowledge can effectively be measured.

CONCLUSION

The discussion contained within this double issue is very important in terms of consolidating and expanding upon our current understandings of what counts as knowledge about language. This collaborative approach to defining knowledge is an essential element of academic discourse and it provides an effective platform upon which to build future practice. Current flawed language assessments, however, stand in the way of real progress in pedagogy. Collaborative approaches to challenging the validity of such tests will help remove this barrier. Additionally, through collaborative design, current problems can be avoided in future assessment development. Language theorists, seasoned educators, students and other stakeholders can work with assessment specialists to help them better understand the constructs they are measuring, and support them as they design tests that better reflect and support pedagogy. In fact, rather than minimizing the expectations for test validity on the basis of construct complexity and the difficulty involved in defining measurable constructs, assessment specialists should recognize the need to engage in collaborative design and should begin building research networks which include teachers, students, language and literacy specialists, curriculum developers and cognitive psychologists. Validity-based research provides both the rationale and the push for collaborative assessment design in language education. The issue is real, the time is now.

REFERENCES

Alberta Education. (2004). *Diploma examination programs.* Retrieved October 24, 2005, from http://ednet.edc.gov.ab.ca/k%5F12/testing/diploma/bulletins/examples_stand/eng30-1_jan2004.pdf

Alberta Teachers' Association. (2005). *Accountability in education.* Retrieved October 24, 2005 from http://www.teachers.ab.ca/NR/rdonlyres/F07B25F6-23C7-4964-95D1-AB18326C7580/0/AccountabilityDiscussionPaper.pdf.

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.

Atwood, M. (2002). *Negotiating with the dead: A writer on writing*. New York: Cambridge University Press.

Britton, J. (1975). *The development of writing abilities (11-18).* London: Macmillan Education.

Calkins, L. (1994). *The art of teaching writing.* Portsmouth, NH: Heinemann.

Coles, W. & Vopat, J. (1985). *What makes writing good: A multi-perspective*. Toronto: DC Heath and Company.

Crawford, L., Helwig, R., & Tindal, G. (2004). Writing performance assessments: How important is extended time? *Journal of Learning Disabilities 37*(2), 132-142.

Didion, J. (1994). Why I write. In G. Muller & A. Crooks (Eds.), *Major modern essayists* (2nd ed.) (pp. 224-230). Englewood Cliffs, NJ: Prentice Hall.

Dillard, A. (1989). *The writing life*. New York: Harper and Row.

Elbow, P. (1981). *Writing with power: Techniques for mastering the writing process*. New York: Oxford University Press.

Emig, J. (1971). *The composing process of twelfth graders* (NCTE research report No. 13). Urbana: NCTE.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8,* 5-16.

Hillocks, G. (1987). Synthesis of research on teaching writing. *Educational Leadership, 44*(8), 71-82.

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

Hillocks, G. (2003). Fighting back: Assessing the assessments. *English Journal, 92,* 63-70.

Kearns, L. (2005). *Literacy testing and "at risk" youth.* Paper presented at the Language and Literacy Researchers of Canada annual meeting, May, 2005, London, Ontario.

King, S. (2000). *On writing: A memoir of the craft.* New York: Scribner.

Lane, S., Parke, C., & Stone, C. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17,* 24-28.

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

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229-258.

Moss, P. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14,* 5-13.

Murphy, S. (2005) *Assessment Allegories.* Paper presented at the Language and Literacy Researchers of Canada annual meeting, May, 2005, London, Ontario.

Murray, D. (1968). *A writer teaches writing*. Boston: Houghton Mifflin.

Murray, D. (1990). *Shoptalk: Learning to write with writers*. Portsmouth, NH: Heineman.

Orwell, G. (1994). Why I write. In G. Muller & A. Crooks (Eds.), *Major Modern Essayists* (2nd ed.) (pp. 90-97). Englewood Cliffs, NJ: Prentice Hall.

See, C. (2002). *Making a literary life*. New York: Random House.

Smith, K. (2005). Missing literacies: An Ontario testing review. Paper presented at the Language and Literacy Researchers of Canada annual meeting, May, 2005, London, Ontario.

Slomp, D. (2005) Serving the test: The influence of high stakes writing assessment on student literacy. Paper presented at the Language and Literacy Researchers of Canada annual meeting, May, 2005, London, Ontario.

Stafford, K. (2003). *The muses among us: Eloquent listening and other pleasures of the writer's craft*. London: University of Georgia Press.

Wiley, D. (1991). Test validity and invalidity reconsidered. In R. Snow & D. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee. J. Cronbach* (pp. 75-107). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Yancey, K. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50,* 483-503.

Zinsser, W. (1988). *Writing to learn*. New York: Harper and Row.

Manuscript received: October 25, 2005
Revision received: December 5, 2005
Accepted: December 12, 2005