Ethical Considerations and Writing Assessment

by David Slomp, University of Lethbridge

In this introductory article, I set the stage for the arguments that follow in each of the contributions to this special issue. First, I critically examine the three pillars of the current Standards--fairness, validity, and reliability--exploring briefly how on their own each concept is insufficient to guiding ethical practice. Then I briefly examine the Standards themselves highlighting their limitations in guiding ethical practice. Finally, I provide a brief introduction to the various dimensions of the theory of ethics we are developing in this special issue.

Large-scale writing assessment has become ubiquitous in North American education. Students at the K-12 level in Canada and the United States are virtually guaranteed to be subjected to any number of large-scale writing assessments at some point in their education. Lazarin's (2014) study of testing in 14 large school districts in seven US states found, for example, that students write as many as 20 (and an average of 10) standardized tests a year. A study conducted by the Council for the Great City Schools, composed of superintendents and school board members from the nation's largest urban school systems, found that students in the 66 sampled districts were required to take an average of 112.3 tests between pre-K and grade 12—a total that does not include diagnostic, school, or teacher developed tests. More specifically, in the 2014-15 school year, students in the 66 urban school districts sat for tests more than 6,570 times (Hart, Casserly, Uzell, Palacios, Corcoran, & Spurgeon, 2015). Faced with increasing oppositions, the Obama administration admitted that testing had gone too far and, as the NY Times reported, acknowledged its role in test proliferation (Zernike, 2015). In its reauthorization of the Elementary and Secondary Education Act of 1965 on January 6, 2015, the Every Child Succeeds Act (S.1177) substantially limits the role of the federal government in education and restores to the states the responsibility for federal test use, with additional support for locally developed assessments.

The stakes associated with these assessments have and will vary from low to extreme, from locally-developed and school-based to standardized and federally-sponsored. Their impacts on students, teachers, and systems of education will vary also. It is within this shifting and contingent environment that the present special issue of the *Journal of Writing Assessment*—that begins to articulate a theory of ethics for the field—is situated.

Some might question the need for a theory of ethics. After all, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) already have defined technical requirements for assessment design and use. Throughout this special issue, however, we argue that technical competence/quality is only one component of ethical practice. Technical quality or feasibility may provide some justification for implementing an assessment practice, but technical feasibility is not equivalent to moral or ethical justification for that practice.

Consider, for example, recent problems with large-scale writing assessments in Alberta, Canada, and Nevada. In both cases the platforms that housed new computer-based literacy tests crashed while students were trying to log in to write their exams. In Alberta, assessment officials made the decision to use regression analysis as a tool for generating a replacement (or fake) grade for all the students who were affected by the crash of the exam platform. Rather than receiving a grade for actual performance on the writing exam these students received a grade that was based on the statistical manipulation of three sets of data: (a) students' scores on their reading comprehension exam; (b) students' school awarded marks; (c) a statistical analysis that compares (a) and (b) against the performance of other students in the province who completed the writing exam.

The decision to generate replacement grades was justified on the basis of three core principles articulated in the *Standards*. First, this approach was *fair* because it attempted to mitigate in as equitable a fashion as possible, the negative impacts for students caused by the exam's crash. Second, this approach is *reliable* "providing the 'best predictor' of how these students would have performed on [the writing exam] if they actually wrote the examination" (Alberta Education, 2015). Third, this approach generates *valid* scores:

> Multiple regression is a method used by Alberta Education to estimate/predict the unknown mark (in this case, part A). It is based on statistical analysis to determine the relationships among three variables (Part A, Part B and School-awarded marks) of unaffected students. These calculated relationships are used to generate the unknown mark for affected students who are requesting a partial exemption. (Alberta Education, 2015)

However, even though the solution is technically feasible and was justified to some degree using arguments related to fairness, reliability, and validity, the ethical questions remain. Is it ethical to generate a proxy grade for students on a high-stakes exam? Is it ethical to use replacement grades on high-stakes exams to determine eligibility for a high school diploma, for high school and university scholarships, and for post-secondary admissions? What are the consequences of this practice for students, teachers, and systems of education? Are these consequences justifiable?

As illustrated by this example, a theory of ethics compels attention beyond the question of technical competence towards broader questions of social consequences. Additionally, the theory of ethics we are developing recognizes that technical standards are themselves social constructions, designed by a community of stakeholders out of a particular perspective and to serve specific purposes. As such, it calls for a critical engagement with those standards, perspectives and purposes.

In this introductory article, I set the stage for the arguments that follow in each of the contributions to this special issue. First, I critically examine the three pillars of the current *Standards*—fairness, validity, and reliability—exploring briefly how on their own each concept is insufficient to guiding ethical practice. Then I briefly examine the *Standards* themselves highlighting their limitations in guiding ethical practice. Finally, I provide a brief introduction to the various dimensions of the theory of ethics we are developing in this special issue.

Sources of Evidence and the *Standards*

Each of the authors of this special issue recognizes no assessment program is neutral. Whether by intention or by fact of their implementation, all such programs have an effect on the individuals and systems to which they connected. Recognizing this fact, the educational community has worked hard over the years to establish conceptual and technical guidelines for managing the influence of assessment programs. *The Standards for Educational and Psychological Measurement* (AERA, APA, & NCME, 2014) defined technical qualities that are essential to the evaluation of tests, testing practices, and test use. These technical standards have evolved over time to reflect advances in research and to address changes in practice and uses of assessments. Historically, technical quality of assessment programs has been defined by these *Standards* in terms of both the concepts of *reliability* and *validity*; more recently the concept of *fairness* has received additional attention. Because of the importance of these key terms and the concepts they suggest, attention to each is warranted.

Reliability as Consistency

Broadly speaking, reliability is concerned with the social and scientific values of dependability, consistency, accuracy and precision (Parkes, 2007). As such, reliability is essentially a facet of the concern for construct validity; low reliability indicates that construct-irrelevant variance is in some way reducing the precision of test scores, and by extension, their dependability. In this way, reliability can also be understood as a form of fairness. An instruments' capacity to produce scores that consistently reflect a precise measurement of the construct enables test users to make fair decisions and inferences about students and their ability. Similar to issues of validity, the demonstration of high degrees of reliability can provide some technical justification for the use of an assessment without addressing deeper ethical questions. Historically, for example, selected response tests that measured writing ability demonstrated high degrees of reliability. In some respect, such instruments also demonstrated certain degrees of validity, yet in the 1960s and 1970s these were largely replaced by tests that measured samples of actual student writing because such tests were seen to be more valid. From the perspective of social consequences, such tests also seemed to better support more effective practices with respect to teaching and learning in schools. In the 1980s and 1990s this shift from selected response test formats for writing toward performance-based assessments of writing saw the development and introduction of portfolio-based writing assessments. The strengths of portfolio assessment are that they enabled test developers and users to capture a broader more complex sample of the writing construct. Their weakness, however, is that they often demonstrated weak measures of reliability. As a consequence, many state assessment programs have abandoned their portfolio assessment programs. Parkes (2007) associated this history with a major problem with reliability theory: The measurement community has for too long conflated the social values underpinning reliability with the narrow set of methods established for measuring the degree to which such values have been captured by a set of test scores.

Validity as Defense

Validity has historically been understood as the primary concern for evaluating the integrity of assessment programs. While the concept itself has evolved over time, it currently refers to the defensibility, and thus to the appropriateness, of our uses and interpretations of assessment results. Huot (2002) has made much of the fact that this current conception of validity places the concept within the domain of rhetoric. In the process of validation, assessment developers and users must construct an argument that defends the uses and interpretations of assessments results. Validity, then, hinges on one's ability to construct an argument. Validity theorists, themselves, have consistently and explicitly narrowed the breadth of such arguments to focus solely on the uses and interpretations of test results. As such, these arguments are framed as technical ones. The questions they are designed to answer is, "On the basis of their technical merits, can we justify the uses and interpretations of these test results?" Historically, answers to this question have been framed in several ways. We can trust them because they (a) have been shown to accurately predict future performance; (b) reflect similar scores achieved on similar parallel measures; and (c) accurately reflect the construct the instrument was designed to measure.

Such questions, however, tend to ignore the broader ethical questions associated with the concept of defensibility. In spite of the fact that (a), (b) and (c) may be true, can we defend the use of assessment results for tests that measure constructs we know little about or for where there is little consensus as to what the construct entails? Can we defend the use of assessments that measure

well the construct they intended to measure, but that are measuring the wrong construct or facets of the construct so narrow that they are irrelevant? Can we defend the use of assessments that only measure the narrow aspects of the construct that can easily be measured by tests?

In Ontario, Canada for example, the Ontario Secondary School Test (OSSLT) was designed to measure the construct of basic literacy. Even if such a test measured the construct of basic literacy perfectly, could it be justified when the very construct "basic literacy" is itself so hotly contended? Can the use of such a test be justified if it fails to capture the broader literacy construct as it is understood both in the Ontario curriculum and in the academic literature? Can the use of such a test be justified if it has been shown to have negative impacts on the broader system of education, on teachers' sense of professionalism, on student self-perception, or on the breadth of the literacy curriculum taught in Ontario schools (Slomp, Corrigan & Sugimoto, 2014)? Validity theory as it has currently been constructed provides no answers to these questions. For this reason, Schendel and O'Neill (1999) argued that valid use is not that same as ethical use. They wrote,

> Although validity is often a part of discussions of assessment, the ethical dimension is often missing. To ensure that our assessment practices are both ethical and valid, we should engage in critical examination of the processes and consequences of asking students to assess their writing as well as the rhetoric we use to talk about [assessment] practices. (p. 200)

Messick's work in the 1980s advocated a return to the ethical aspect of validity by calling for assessment developers and users to examine both the actual and potential consequences of assessment design and implementation (Mike, 2013). Yet, while Messick's move to make construct validity the central concern in validity theory has been widely accepted within the field of educational measurement, his simultaneous move to fuse concerns for construct validity with concerns for the consequences of test use have not received the same level of acceptance. As is the case with reliability, validity can only take us so far in making decisions about the ethical use of assessments.

Fairness as Validity

The most recent version of the *Standards* marks a radical step forward from earlier editions by elevating the concept of fairness to a level equal to that of the concern for both reliability and validity. A concern for fairness, however, has been an overt goal of most large scale assessment programs dating as far back as the Imperial Chinese examination program. However, because assessment always involves a power imbalance between those who ask questions and those who are required to answer them, Spolsky (2014) argued that other unstated purposes have often been the true drivers of such assessment programs: In imperial China assessment was used to control the less privileged, and to select among them; in 19th Century England, the civil service examination was designed to replicate the social order of the day; in the 1950's Australia's immigration test was designed to control immigration patterns for certain ethnic groups; in the 1960s the TOEFL was also used to "control the immigration loophole" (p. 1575). Spolsky's history makes clear that fairness, understood as a technical concern, should not be equated with ethical practice. The Imperial Chinese civil service exam may have been designed to select as fairly as possible candidates for the civil service while concurrently operating as an instrument of social control. In current times, large-scale high-stakes writing assessments may be designed to reflect principles of fairness for individual students while simultaneously being employed to both control and shape education systems. In such cases, these assessment programs may be technically sound while also being morally debatable. As such, the definition of fairness in the *Standards* (AERA, APA, & NCME, 2014)—a fair test minimizes variance that "would compromise the validity of scores for some individuals" (p. 219)—seems quite beside the point both in its self-referential solipsism and silence on consequence.

Limitations of the *Standards*

Each of these concepts have been defined and updated repeatedly in the *Standards*. The *Standards* themselves have been created to guide assessment design and use and will continue to play an important role in educational measurement in general and writing assessment in particular.

The *Standards*, however, are nevertheless insufficient for guiding ethical decision making: They reflect a narrow epistemological, ontological and axiological standpoint; they focus narrowly on intended uses and interpretations of test scores; and they handle key technical issues such as validity, reliability, and fairness as siloed concepts. An important flaw in the *Standards* is that they are designed to reflect the perspectives and interests of the dominant stakeholder group—those who design and use large-scale assessments (Maul, 2014, p. 40)—while simultaneously excluding the perspectives of classroom teachers (Plake & Wise, 2014). As a result, they fail to attend to the broader social consequences that Messick advocated attention to.

While the *Standards* acknowledged reliability, validity, and fairness are related concepts, it treats them independently of one another while at the same time calling on test users and developers to make integrated judgments regarding assessment design and use. Unfortunately, the *Standards* provide only the vaguest of guidance on how such integrated judgments should be structured:

> [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 that is well supported and defendable (AERA, APA, & NCME, 2014, p. 22).

What the field requires is a more cohesive, integrated framework that provides more concrete guidance for assessment design and use.

Taken as a whole, The *Standards* pay little attention to a systems-level perspective on the role of assessment in education (Diaz-Bilello et al., 2014). In the United States educational policies such as No Child Left Behind (NCLB) and Race to the Top (RTTT) have created an environment in which testing has become an apparatus within larger systems of accountability. This phenomenon is not unique to the American context, as systems of education around the globe are increasingly administered within rigid accountability frameworks. Within such accountability systems, technical quality of testing instruments becomes increasingly important. The *Standards* play an important role in this respect. However, technical quality in itself is insufficient; accountability systems themselves need to be critically evaluated, their impact on the systems over which they have been imposed need to be rigorously evaluated, and the responsibilities of both those who design these systems and those who enable their use—both test users and test designers—need to be defined and enforced. The sub-prime credit crisis at the turn of the current century provided ample examples of how flaws in accountability mechanisms can have catastrophic consequences for the systems over which they have been imposed.

A Role for Ethics

As is the case in the *Standards*, fairness has remained wedded to instrumental concerns in contemporary measurement theory. The concerns are explicitly evidenced in the 2010 issue of *Language Testing* in which Xi situated fairness within the framework of validity: “Fairness is characterized as comparable validity for relevant groups that can be identified. The fairness argument consists of a series of rebuttals that may challenge the comparability of score-based decisions and consequences for sub-groups” (p. 167). Likewise, The *Standards*’ treatment of fairness remains rather cosmetic, essentially treating fairness as a subset of validity. For example, the current *Standards* limited their concern for subgroup difference to the issue of construct irrelevant variance and construct underrepresentation. Broader issues related to cultural bias—such as subgroup differences being related to undemonstrated assumptions about students rather than from reflective latent variable models validated under field-test conditions—are not taken up in the *Standards*. For reasons such as this, ethicists have made the point that technical competence is not synonymous with ethical use. While necessary, technical competence is an insufficient justification for use; simply because something is technically feasible does not make it morally or ethically justifiable. Indeed, focusing on the technical aspects alone holds the danger of technological determinism.

As is the case with bias, fairness in educational measurement has primarily been addressed through comparing items and test performance in different identifiable groups. Camili (2006) refered to these techniques as the structural analysis of bias (including use of such models as differential item functioning) and external evidence of bias (including regression models to identify differential prediction). Our goal in this special issue was to interrogate fairness under equally rigorous philosophical frameworks, paying special attention to current writing assessment frameworks that call for recognizing the social dimensions of assessment: local considerations, community-based assessment, and the effects of assessment. Yet this philosophical approach raises a critical question: How can we further an agenda for fairness if we cannot identify—and challenge—the philosophical tradition from which it arises?

Of the three guiding principles—validity, reliability, and fairness—fairness, with its attention to impacts of assessment practices on individuals, touches most closely on the need for new practices informed by moral philosophy. While definitions of ethical behavior date from antiquity, a contemporary definition of ethics by James Rachels (2012) in the *Elements of Moral Philosophy* affords an initial context to situate fairness within a broad philosophical realm: agentic. Rachels framed his definition in terms of the conscious moral agent as

> someone who is concerned impartially with the interests of everyone affected by what he or she does; who carefully sifts facts and examines their implications; who accepts principles of conduct only after scrutinizing them to make sure they are sound; who is willing to ‘listen to reason’ even when it means that his or her earlier convictions may have to be revised; and who, finally, is willing to act of the results of this deliberation. (p. 11)

While we may argue that Rachel’s definition is decidedly western in its reliance on reason and careful sifting of facts as a path toward decision-making, our line of inquiry begins with this tradition because it a toehold into the steep cliff upon which measurement theories of fairness have been based. From Socrates to MacIntyre, a distinct set of qualities—emphasis on systematic reasoning, commitment to principled action, and concern for others—remains at the heart of western orientations toward how we might best live. Indeed, for Rawls (1999, 2001) justice as fairness became central to his theory because it allowed both emphasis on obligations and attention to the individual.

Narrowing further, fairness (obligatory aims in pursuit of equality of opportunity) is taken to be a distinct line of ethical inquiry (varied actions in pursuit of socially constructed concepts of the good). Because it is beyond the scope of this special issue to outline a comprehensive agenda common to each article, the special issue is best understood through identification of facets of fairness associated with writing assessment. By extension, articles in this special issue include attention to the following:

- *Sociocultural perspectives* on the origin of traditions, with attendant acknowledgement of the limits of practices redolent of colonialism and capitalism;
- *Access* to educational structures that are associated with literacy;
- *Opportunity to learn* as an often forgotten aim of assessment and a controlling factor in allocation of instructional resources;
- *Maximum construct representation* that is clearly articulated in advance of the assessment and neither implicit not derived through post-hoc methods;
- *Disaggregation of data* so score interpretation and use can be clearly understood for all groups and each individual within those groups;
- *Justice* as a principle of fairness so opportunities do not merely exist but, rather, so each individual has a fair chance to secure such opportunities

While our authors define unique implications and applications of this definition, each holds firm belief in the following facets of the theory: the significance of the specific institutional site; the relevance of social sociocultural perspective; the importance of advancing opportunity to learn for both groups and individuals; the need for robust construct representation in terms of assessment advantage for all students; the relevance of refusing to fix pre-established definitions of the least advantaged; the need to secure resource allocation for those disadvantaged by the assessment; and the use of varied quantitative and qualitative techniques to ensure an actionable agenda for fairness.

Despite the comprehensive treatment of the authors of the special issue, each author agrees that significant questions remain for readers:

- Is fairness reactive or proactive?
- Where does fairness intersect with transformation and care?
- How can fairness account for what is unwitting or invisible in daily practice?
- How do we identify least advantaged when often such groups are not easily identifiable?
- Following identification, what is the role of agency when discussions of the least advantaged occur?
- What actions can or should lie within the reach of fairness?
- Because it is not solely a technical or measurement term, who ultimately owns fairness?
- What is to be done when the very cultural frame in which we work, one often associated with meritocracy, remorselessly denies working toward the benefit of the least advantaged?
- How can non-western traditions be brought to bear on fairness in writing assessment?

Ethics and Writing Assessment: Necessity and Sufficiency

Given the both the necessity yet insufficiency of foundational design principles of fairness, validity, and reliability in guiding ethical decision-making, a new unifying framework is needed; one that advances broader ethical concerns in the design, implementation, and use of tests. To this end, we are proposing a theory of ethics for the field of writing assessment, one that advances such a framework toward new conceptualizations that better serve students. Such a theory should assist all stakeholders in the assessment process in more thoroughly addressing questions regarding the moral aspects of assessment use. As such, we believe a theory of ethics for writing assessment must:

- Be the driving concern of educational stakeholders—the primary referential frame that conceptualizes instruction and assessment in terms of each other in ontological, epistemological, and axiological perspectives.
- Explore issues related to reliability and validity from multiple ontological and epistemic and axiological stakeholder perspectives concerned with fairness, thereby offering an overall referential frame on what constitutes writing assessment that is robust enough to justify various uses of scores.
- Have an ecological orientation; one that pays attention to the role assessment plays both within broader systems of education and within society as a whole. It needs to account for how assessments shape systems of education, and how they impact stakeholders within those systems.
- Provide a unifying function, one that draws together concerns for validity, reliability, and fairness, and an advancing function, one that ties these concerns to ethical decision-making. It must account for the perspectives and experiences of key stakeholders within the measurement process.
- Have value for a range of assessment contexts, both large scale, standardized testing and locally-developed, site-based assessments.
- Hold test-users to actionable standards of ethical practices, and require assessment developers—whether site-specific or large scale—not allow themselves to become complicit in the unethical use of their tests (either by refusing to bid on RFPs that

require they violate their standards, or by failing to publicly call attention to unethical uses of tests they have developed).

We offer this theory in the spirit that Gloria J. Ladson-Billings expressed in her lecture following her receipt of the 2015 Social Justice in Education Award, when she stated she wanted to "trouble" the term social justice. She asked her audience to participate in a fundamental rethinking of our past and our work as human beings. Similarly, in this special issue we challenge our colleagues to critically rethink the historical theories and frames that have shaped our field, and to reimagine future possibilities. Social justice, she held, is not a concept expansive enough to confront the injustice that holds a deadly grip on our society. While we will surely differ in our concepts of moral philosophy, ethics, and fairness, our aim is at one with hers in the pursuit of justice for our students.

References

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

Alberta Education (2015). Alberta Education. Retrieved from http://schools.cbe.ab.ca/b870/Publications/CBE%20ELA%2030-1%20and%2030-2%20Diploma%20Exam%20Exemption%20FAQ%20Document.pdf

Diaz-Bilello, E., Patelis, T., Marion, S., Hall, E., Betebenner, D., & Gong, B. (2014). Are the *Standards for Educational and Psychological Testing* relevant to state and local assessment programs? *Educational Measurement: Issues and Practice, 33*(4), 16–18.

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., Spurgeon, L. (2015). *Student testing in America's Great City Schools: An inventory and preliminary analysis*. Washington, DC: Council of Great City Schools. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

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

Ladson-Billings, G. J. (2015, April). *Justice...Just Justice.* Social Justice in Education Award lecture presented at the American Educational Research Association annual meeting, Chicago, IL. Retrieved from https://www.youtube.com/watch?v=ofB_t1oTYhI

Lazarin, M. (2014). *Testing overload in America's schools*. Washington DC: Center for American Progress.

Maul, A. (2014). Justification is not truth, and testing is not measurement: Understanding the purpose and limitations of the *Standards*. *Educational Measurement: Issues and Practice, 33*(4), 39–41.

Mike, G. (2013). Towards an ethics of writing placement. *CEA Critic, 75*(1), 51–65.

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice, 26*(4), 2–10.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing*? *Educational Measurement: Issues and Practice, 33*(4), 4–12.

Rachels J. (2012). *The elements of moral philosophy* (7th Ed.). Singapore: McGraw-Hill International Edition.

Rawls, J. (1999). *A theory of justice* (Rev. ed). Cambridge, MA: Cambridge University Press. (Original work published 1971)

Rawls, J. (2001). *Justice as fairness: A restatement* (R. Kelly, Ed.). Cambridge, MA: Harvard University Press.

Schendel, E., & O'Neill, P. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing Writing, 6* (2), 199–227.

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.

Spolsky, B. (2014). The Influence of ethics in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3) (pp. 1571-1585). Oxford, UK: John Wiley and Sons.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.

Zerinike, K. (2015, October 24). Obama administration calls for limits on testing in schools. *The New York Times*. Retrieved from http://www.nytimes.com/2015/10/25/us/obama-administration-calls-for-limits-on-testing-in-schools.html?_r=0