INTERNATIONAL LITERACY ASSOCIATION
JOURNAL OF ADOLESCENT & ADULT LITERACY

Sex, Finance and Literacy Assessment

Journal:	*Journal of Adolescent & Adult Literacy*
Manuscript ID	JAAL-2020-02-0029.R1
Manuscript Type:	Department
Extended Keywords: Please select extended keywords to make your paper more searchable online:	Assessment, Standardized < Assessment, To inform instruction, as inquiry < Assessment
Level of Learners: Please select as many Level of Learners as apply :	2-Childhood, 3-Early adolescence, 4-Adolescence, 5-College/university students
Keywords: Please select keywords for peer review:	Assessment, Writing, Language learners
Abstract:	Use the standing abstract.

Sex, Finance and Literacy Assessment

As literacy educators we often handle discussions about assessment like we treat chats between spouses about finance, or talks between parents and children about sex. It's complicated. We know how we feel about it, what we think about it, how we handle it, but we often aren't always confident in what we know about it. It's contentious. There are many view points and perspectives on it, and lots of experts who can tell us what to think about it and how best to manage it. It doesn't help that experts don't always agree, or that their opinions and advice often contradict each other.

Further complicating matters, we often aren't sure about where others stand or what perspectives they hold. Given all the complexity and uncertainty around the topic, it sometimes seems best to simply avoid the conversation. But, like issues related to sexuality and finance, issues surrounding assessment are, for literacy educators, a fact of everyday life.

If you work in the field of literacy education, inevitably you deal with assessment on a daily basis. In our classrooms, literacy educators create assessments to manage student behavior and to monitor and support student learning. Additionally, we inevitably deal regularly with external assessments that are imposed on us, our classrooms, and our students. Literacy is the one assessment topic found in virtually every state, provincial, national or international assessment program.

When editors Kathleen Hinchman and Kelly Chandler-Olcott invited me to edit a department on literacy assessment, they asked me to create a column that would help spur conversations about assessment in *JAAL* and amongst members of the International Literacy Association. In accepting this role, I couldn't promise them the wisdom and eloquence of either sexologist Dr. Ruth or personal finance guru Dave Ramsey.

At the very least I hope to offer a column that is direct, straightforward, and informed; one that helps to make the necessary conversations easier, more confident, and more impactful. With this goal in mind I have invited some leading literacy assessment researchers to co-author a series of columns that each explore the challenges involved in doing literacy assessment well. I hope these columns will provide the readers of JAAL with some of the tools needed to design and advocate for quality literacy assessments for your classrooms.

I bring to this column more than 20 years of experience as an English language arts teacher and literacy assessment researcher. I've taught literacy to students in junior and senior high schools, in adult academic upgrading (high school equivalence) and ESL community college classrooms, and in university writing and literacy education courses. I now serve as an associate professor of literacy assessment at the University of Lethbridge in Alberta, Canada where I co-direct a Masters of Education program in Curriculum and Assessment. I currently co-edit the international journal *Assessing Writing*, a position that provides me with a front row seat on the latest advances from around the world in the field of writing assessment.

Framing Difficult Conversations

Difficult conversations rarely go well when parties to them do not share similar perspectives, frames, or values on the topic at hand. Problems arise amongst spouses, for example, when the free-spender and the tight-wad don't come to shared values around money.

Similarly literacy educators frame assessment as a tool for supporting teaching and learning, while the general public, politicians, and measurement specialists see large-scale assessment as a tool for surveillance and accountability (Ydesen & Bomholt, 2019).

As those closest to the action, literacy educators are often the first to see problems with large-scale assessments. However, when they raise their concerns these are seldom heard and

rarely acted upon. In my 20 years of advocating for assessment reform, I've observed that literacy educators and decision-makers often talk past each other because they frame their concerns so differently.

Fortunately, the conceptual frames—validity, fairness, and reliability—that shape the measurement community's thinking about quality assessment can also be used by literacy educators to advocate for quality assessments that support teaching and learning.

Validity. When we validate an assessment, we test how well it serves each purpose for which it was designed. For this reason, the first step in the validation process is to determine why an assessment is being used. Questions to be answered at this stage include: (a) What purpose is this assessment designed to serve? (b) What inferences about student learning or program quality will be drawn from student performance data? (c) What decisions will be made based on those inferences?

Each proposed inference and each proposed decision must then be tested. In the resulting validity argument a series of claims are stated, and evidence is collected to test those claims (see Table 1).

Each set of claims must be supported by evidence. If one claim fails, the entire validity argument fails and the assessment must either be revised or discontinued (Kane, 2013).

Literacy educators are well positioned to observe first hand when any of these claims have failed. For example, as professionals who understand the diversity and complexity of literacy, teachers often know intuitively when an assessment is based on a limited literacy construct. For example, many large-scale writing assessments reward polished, first-draft writing, an assessment design that excludes opportunities for revision. As a consequence, students learn to see writing as a single draft process, a perspective that undermines their long-

term growth, development and resilience as writers. Similarly, large-scale reading assessments almost always position reading as an efferent activity while ignoring the aesthetic dimension. Aesthetic experiences with reading, however, are the foundation to life-long enjoyment with text. It is little wonder that for many students the love of reading diminishes as they progress through school.

From a validity standpoint, these design choices undermine the inferences and decisions that can be made from these assessments while also promoting negative long-term outcomes for students.

Fairness. At its core, fairness is concerned with equity. In contemporary assessment theory the concern for equity is articulated both in terms of aspiration—Equity of Opportunity to Demonstrate Learning, Opportunity to Learn—and melioration—Absence of Bias, and Disparate Impact—(see Table 2). When advocating for fairness our goal is to ensure that an assessment does not systematically disadvantage one of group of students compared to others (Nisbet, & Shaw, 2019).

Once an afterthought in measurement theory, fairness has been elevated in recent years to a position of primary importance (Elliot, 2015). This shift reflects the increasing diversity, and complexity, of our classrooms and the increased ambition of our pedagogies as we commit to ensuring all students are able to achieve and succeed. Reflecting this commitment, our assessments must ensure that all students have an equal opportunity to demonstrate their abilities relative to the constructs they are measuring.

As with issues of validity, classroom educators, because they engage with and monitor student performance on a day-to-day basis, are best positioned to recognize issues of fairness.

The first step in examining fairness is to analyze student performance disaggregated by populations of interest. Does one group of test-takers unexpectedly under-perform another group? If the answer to this questions is yes, a closer investigation is needed to see if issues of bias or inequity are impacting assessment results.

A second step is to analyze test items to see if they require cultural knowledge that is irrelevant to the task. This analysis should be followed with an examination of how test-takers interpret and respond to test items to see if these items require different knowledge and skills from one group than they require from another. For example, fairness issues can arise when a writing task requires irrelevant cultural knowledge to complete. Recent immigrants might possess the writing skills being measured but can't demonstrate their abilities because they lack access to the required but irrelevant cultural knowledge.

Opportunity to Learn is the broadest fairness issue (Moss, Pullin, Gee, Haertel, & Young, 2008). Investigations into this issue examine the degree to which all populations have access to basic conditions (social, emotional, economic) needed to thrive.

Reliability. In many respects, reliability—a concern for the social values of consistency, dependability, and replicability—is a precondition for both validity and fairness (Parkes, 2007, 2013).

Historically, large-scale assessments have been designed to achieve high degrees of reliability. Unfortunately, this focus on reliability has often come at the expense of validity, as it is easier to consistently measure a narrow more limited construct than it is to measure a more complex construct (Slomp & Fuite, 2004). Teachers can examine issues of construct misrepresentation brought on by design choices that privilege reliability over validity by

examining test items for issues of systematic error, and by examining both test-taker and raters' cognitive processes to determine sources of variance in scores (see Table 3).

Beginning and Ending with Consequences

Similar to entering into conversations about sexuality and finance, it is important to begin conversations about literacy assessment by recognizing the consequential nature of what is being discussed. In fact, it is often a concern for consequences in the first place that drives the need for the difficult and awkward conversations that follow.

Assessments are always consequential. Every assessment carries with it the potential to positively or negatively impact individuals, populations, teachers, schools, or entire systems of education. For example, when large-scale essay exams only measure polished first-draft writing, they teach students to see writing as a single draft exercise, they coerce teachers into focusing their instruction on drafting rather than revising skills, and they create conditions within entire systems of education where risk-taking and exploration are disassociated from the process of writing (Slomp, Graves, Broad, 2014).

Literacy educators shouldn't shy away from the difficult conversations. Rather, understanding the values that frame quality assessment, we need to advocate for practices that enable us and our students to thrive.

References

Elliot, N. (2015). Review essay: Validation: The pursuit. *College Composition and Communication, 66*(4), 668-685.

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

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge University Press.

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612-629.

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

Parkes, J. (2013). Reliability in classroom assessment. *SAGE handbook of research on classroom assessment*, 107-124.

Slomp, D. H., & Fuite, J. (2004). Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing*, *9*(3), 190-207.

Slomp, D. H., 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.

Ydesen, C., & Bomholt, A. (2019). Accountability implications of the OECD's economistic approach to education: A historical case analysis. *Journal of Educational Change*, 1-21.

Table 1

Structure of a Validity Argument

Inferential Chain	Claim to be Tested	Questions to Answered	Evidence to be Collected
Construct Definition*	Assessment is based on a robust, research informed construct.	Does a detailed, research informed description of the literacy construct provide a foundation for the assessment's design?	Systematic review of the literature describing knowledge, skills, and dispositions related to literacy (or aspect of literacy) being assessed. Construct map or model based on systematic review.
Construct Sample	Assessment captures a balanced and representative sample of the construct being measured.	Does the assessment blueprint demonstrate that the assessment is designed to capture both breadth and balance across the construct elements? Does the scoring criteria reflect appropriate breadth and balance across construct elements?	Assessment blueprint that demonstrates links between assessment features, scoring criteria, and construct elements. Expert review of assessment blueprint.
Scoring Inference	Each student's scores *accurately reflect* their performance on the construct being measured.	Do the environment in which the assessment takes place, the tasks to be completed, and the scoring criteria and processes used accurately capture each student's ability with respect to the construct sample being measured?	Evidence of consistency in scoring between raters. Evidence of consistency in each scorer's judgements. Think aloud data from test-takers that examines alignment between cognitive processes and scores received.
Generalization Inference	Results from the assessment represent what a test-taker would be expected to obtain over multiple similar tasks completed in multiple assessment sessions.	Is student performance on this assessment consistent with their performance on other assessments of the same construct?	Data comparing assessment scores against scores on other similar measures or on variations of the same assessment administered several times.
Extrapolation Inference	Results from the assessment represent how a test-taker would be expected to perform in non-testing (that is real-life) contexts.	Does the environment, range of tasks, and breadth of scoring criteria enable inferences to be made about test-takers' literacy abilities in real-life contexts?	Evidence that assessment scores predict performance in non-testing situations (e.g. test-scores correlate with future GPA; literacy test scores reflect literacies used outside of school).
Decision Inference	The assessment contributes to decisions that are justifiable, equitable and based on sufficient and useful evidence.	Does the evidence collected provide confidence that assessment driven decisions are justified?	Integrated review of evidence collected for each of the preceding claims.

*A construct definition is a detailed description of the knowledge, skills, and dispositions that describe the phenomenon being assessed

Table 2

Key Elements of Fairness

Concern	Questions to ask	Evidence
Opportunity to Learn	Do students have equitable access to resources needed to learn what was being assessed? Does each student's ecological context provide equitable opportunity to learn?	Transparency with respect to what is being assessed. Alignment between assessment, curriculum, and instruction.
Equity of Opportunity to Demonstrate Learning	Do students have equitable access to the resources and assessment conditions needed to complete the assessment task?	Opportunity to acquire prerequisite knowledge and skills needed to complete the assessment. Equitable funding to support all groups of learners.
Absence of Bias	Do assessments privilege construct-irrelevant skills or knowledge that might differentially disadvantage some students over others?	Disaggregated reporting of performance data Differential item functioning Test-takers' response processes
Disparate Impact	Do some groups of students (e.g. gender, cultural, ethnic, religious) systematically underperform other groups of students?	

Table 3

Elements of Reliability

Reliability Dimension	Definition	Question	Evidence
Systematic Error	Consistent, repeatable error caused by flaws in the assessment design or testing conditions.	Is the assessment free from flaws in the design that create conditions for repeatable error in student performance data?	Expert review of assessment design
Random Error	Imprecision in measurement based on ecological or intra-personal factors	Was the assessment impacted by environmental factors that impacted student performance in unpredictable ways?	Student response process data
Inter-rater Reliability	Consistency of measurement across raters	To what extent do differences in rater perspective, bias, and/or severity impact agreement between raters?	Rater cognitive process data
Intra-rater Reliability	Consistency of measurement by a single rater	To what extent do rater judgements fluctuate across samples being scored, when controlled for quality of samples being scored?	Statistical comparisons of scores by rater and/or scoring criteria.