Assessing Writing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Assessing Writing

Book review

Very Like a Whale: The Assessment of Writing Programs, E.M. White, N. Elliot, I. Peckham. Utah State University Press, Logan (2015). 202 pp., ISBN: 978-0-87421-985-2

Very Like a Whale: The Assessment of Writing Programs (Whale) is Ed White, Norbert Elliot, and Irv Peckham's most recent offering to the writing assessment community. *Whale* captures effectively the complexity involved in assessing writing programs. It challenges writing program administrators, writing instructors, and those responsible for assessing writing to design and execute high quality writing assessment programs that model the constructs associated with writing development for students within their unique institutional ecologies (p. 7).

To begin, a note about the audience, focus, and purpose of this book. White, Elliot, and Peckham have written this book for teachers of writing and for writing researchers. The book focuses on the assessment of writing programs at postsecondary institutions in the United States. The authors' goal is to communicate assessment concepts to professionals in the fields of composition, rhetoric and writing.

Others have written reviews of this book for American postsecondary writing professionals (Melzer, 2015; O'Neill, 2015). The audience for *ASW*, however, is broader and more diverse than this group and so the question I have used to frame this review is "What does *Whale* offer to the rest of us—readers who either are not American or WPAs but who have an interest in the field of writing assessment?"

1. Core idea: humbled by complexity

Faced with the task of designing assessment programs, the tendency, too often, has been to look for quick solutions, for easy answers, for actionable insights, for templates of assessment programs that can readily be adopted. This book, does not offer that. Instead, it challenges its readers to step back, to consider the enormous challenge of the undertaking, and to be humbled by it. In what is perhaps the most important passage of this text, White, Elliot, and Peckham write:

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

This passage evokes Stephen Jay Gould's (1996) *The Mismeasure of Man*, and the sense of hubris associated with early attempts to understand the construct of intelligence. Gould illustrates how that sense of certainty led at times to indefensible interpretations and unjustifiable uses of assessment data. Research into the social consequences of both large-scale and classroom assessments of writing suggests that our field, too, has much reason for humility (Hillocks, 2002; Slomp, Corrigan, & Sugimoto, 2014). Indeed, this is an important frame within which to consider what *Whale* has to offer its readers.

The above quote from *Whale* also points to another significant accomplishment of this book: its emphasis on the importance of teacher knowledge in assessment. For decades now, *The Standards of Educational and Psychological Testing* have downplayed and even ignored teacher perspectives on assessment privileging instead expert assessment knowledge (Maul, 2014; Plake & Wise, 2014). The imposition on teachers of external assessments is a natural extension of this stance. In Alberta, Canada, where I teach, the argument for increased frequency of such assessment has been based on the poor quality of teacher-made classroom assessments (Marcellus, 2014). In *Whale*, however, White, Elliot, and Peckham demonstrate that the answer to poor quality classroom assessments is not more external assessment; the answer is building teacher capacity for creating high quality locally developed assessments. This is an important response to a longstanding concern.

http://dx.doi.org/10.1016/j.asw.2016.01.002
1075-2935/

2. Overview: elements of purposeful design

The introduction to *Whale* establishes two key principles that guide the authors' view of writing program assessment. First, in an era of accountability, those who design writing programs must take a proactive stance. Anticipating the demands for accountability when they design their programs, they must build into them the opportunities to collect the data needed for quality control and accountability purposes. Linked to this idea, they emphasize the importance of localism. High quality locally developed assessments, tailored to the jurisdictions, institutions, or programs to which they are linked can provide data needed to monitor those programs while at the same time, enabling comparisons across programs—mindful of the circumstances that shape them—that enable us to search for resonances across diverse contexts. As such, *Whale* advances an approach to program assessment that emphasizes local development, collaboration and accountability.

Chapter 1—Trends—opens with a brief and narrow review of the history of writing program evaluation and accreditation in the United States. This review demonstrates that over the past forty years there has been an increasing sophistication in our understanding of the construct of writing ability, a deepening of our recognition of the complexity of this construct, and a growing awareness of the challenges involved in both capturing such a construct and in measuring a writing program's capacity to foster student development with respect to that construct. Drawing on this history, the authors identify three tropes for conceptualizing the assessment of writing programs: (1) Writing program assessment as a genre: a socially situated tool useful for understanding the cultural rationality of our field and for indexing and understanding our profession. Assessments, after all, are often the most concrete manifestations of what we value about writing within our institutions, our programs, our classrooms. (2) Writing program assessment as a form of construct modeling. The complexity of the writing construct demands that we develop a clear and detailed understanding of the construct we are trying to measure, and that we map a network of assessments onto that construct. The complexity of the construct, White, Elliot, and Peckham argue, emphasizes the need for a network of assessments rather than single assessments. (3) Writing program assessment as ecological study, that is, a mechanism for the examination of the interrelation between evolving organisms and their environment (physical, biotic, interspecies and intraspecies). An ecological lens highlights, most of all, the enormous complexity of the task—we are not measuring simply the traits of student produced text; we are examining those traits within the physical, social, and intrapersonal contexts in which writing ability is being developed and employed.

Chapter 2—Lessons—explores three instrumental case studies of writing program assessments at Louisiana State University, New Jersey Institute of Technology, and the California State University system. The case studies emphasized the importance of robust construct representation and the need for procedures to ensure the validity and validation of these assessment programs. These case studies also demonstrate how such processes lead to an ongoing critical examination and refinement of these writing programs. Drawing on these case studies the authors identify principles for the design of writing program assessments:

- They should emphasize consequences, focusing on the enhancement of student learning and the improvement and sustainability of program quality.
- They should be informed by current scholarship and research.
- They should collect and report data in multiple forms and for multiple stakeholder audiences in a manner that supports evidence-based decision making.
- They should be viewed as a process through which these goals are achieved rather than simply as a product that must be produced.

Chapter 3—Foundations—introduces readers to the concepts of validation, validity evidence, and construct span. The authors emphasize the importance of construct mapping: a process of clearly articulating and defining the construct one is assessing. The hypothetical construct map they provide in this chapter illustrates the enormous complexity of the construct. The authors argue—illustrating with an extended example—that the construct map provides a mechanism for longitudinal program design, course development, and both classroom and program assessment design. Emphasizing the importance of this process, they state,

> We feel certain that, in the near future, it will be difficult for a writing program to be assessed in a valid fashion—perhaps even for it to exist—unless such a figure is available for all stakeholders as a representation, subject to revision and improvement, of the construct that consumes so much time, attention, and resources (p. 76).

This insight is drawn from the White, Elliot, and Peckham's rich understanding of validity theory and its history. Contextualizing this emphasis on construct representation, they pull the thread from Cronbach and Meehl (1955), to Messick (1989) and finally to Kane (2006) emphasizing their insistence on construct representation in the process of designing and validating assessments.

Emphasizing the consequential dimension of validity, the authors, present eleven categories of evidence that can be used to guide the design and validation of writing assessment programs. These categories include: consequence, aim, construct modeling, scoring, disaggregation, generalization, response processes, extrapolation, theory-based interpretation, cost effectiveness, and sustainability. The benefit of this model is that it expands the four categories of validity evidence suggested by

Please cite this article in press as: Slomp, D.H. Book review. *Assessing Writing* (2016), http://dx.doi.org/10.1016/j.asw.2016.01.002

Kane (2006) thereby providing writing assessment scholars with a framework for designing programs of validation research that rely on more than simply post hoc analysis of test results.

Chapter 4—Measurement—discusses measurement concepts that are of importance to the field of writing program assessment: descriptive statistics, sampling-plan design, null hypothesis significance testing, correlation analysis, tests of statistically significant difference, regression analysis, and effect-size determination. This is a difficult but useful chapter, especially for readers who are not trained in measurement. Anticipating resistance from their main audience of writing program administrators, the authors open the chapter by arguing, "the use of descriptive and inferential statistics is more effective in administrative meetings than the rhetoric often employed" (p. 114) and that as a consequence those who advocate for writing programs need to be equipped with the conceptual and empirical tools needed to leverage those methods. I am sympathetic to this argument, though I recognize the concern that such a stance may privilege quantitative data and its associated assumptions and modes of reasoning above those of a qualitative stance. If there is a weakness in this chapter, it is that the authors do not grapple with this tension; in fact, in their appeal to reject value dualisms, they seem to minimize this concern. The differences in stance between quantitative and qualitative reasoning must not be so lightly glossed over because doing so can lead to muddied thinking, poor conclusions, and dangerous actions.

Throughout the chapter, however, the authors demonstrate that they are only too aware of these issues. In their discussion of null hypothesis testing, for example, they demonstrate why correlation coefficients for portfolio assessments must necessarily be lower than those for timed impromptu essays—one cannot expect the same levels of inter-rater reliability for assessments of complex constructs that one expects from those that measure simple (or simplified) constructs. Similarly, in their discussion of regression analysis they demonstrate how the logic of regression for evaluating the success of placement testing necessarily must be different from the logic involved in other forms of predictive validation—if placement testing is successful, students will learn and so post-placement scores may not follow expected patterns. Concluding this example, they make the point, "in making causal inference, the use of empirical information alone can never be valid unless it is accompanied by sound causal logic" (p. 129). This is an important assertion: it balances the authors' call to take up quantitative forms of reasoning with a caution not to do so slavishly.

Chapter 5—Design—opens with the warning that in the current age of accountability, programs must justify their existence to external stakeholders. In this context, the authors argue, high quality assessments are imperative, because those "who have learned to take these matters in stride, turning such demands to [their] program's advantage, will prosper" (p. 145). This position very much reflects the reality of postsecondary education in the United States. In Canada, where I work, the pressures of accountability and competition for government funding are not nearly so severe. However, at all postsecondary institutions competition for resources is a fact of life, and so the caution is an important one. Though White, Elliot, and Peckham themselves do not make the link, this warning is particularly potent for those who work in the K-12 system. Classroom teachers regularly face the pressure to prepare students for externally imposed assessments as a means to justify their employment. The lack of attention in teacher education programs to cultivating teacher capacity for designing high quality assessments (resulting too often in poorly designed classroom-based assessments) speaks to the need for improving classroom assessment capacity (Campbell, 2013). Whether one embraces high quality locally developed assessments because of external pressures, as a tool for activism, or out of a motivation to develop the strongest program possible, the lessons offered in *Whale* are useful for achieving those ends.

In Chapter 5 the authors present their concept, *Design For Assessment (DFA)*, as a model for designing high quality writing assessment programs. Claiming, "the new territory before us will not be hospitable to the ad hoc, intuitive reasoning that all too often continues to inform assessment design" (p. 153), they call for a more purposeful and thoughtful approach to assessment design. DFA begins with a consideration of the aims and consequences of assessment design and use. It requires critical consideration of validation arguments, experimental designs and operational processes required of the assessment program prior to its implementation and use. With its emphasis on attention to consequences, DFA acts upon the need for robust construct articulation and modeling, it responds to the importance of situated knowledge, it frames writing assessment as research (not technology), and highlights the need for methods of documentation that enable integration of quantitative, qualitative, and mixed data. DFA enacts the understanding that accountability should be understood as a form of opportunity, and that the call for sustainability should be seen as an invitation to extend our influence. In short, DFA envisions a process that creates opportunities for transformation of writing programs. It is a hopeful vision—complicated, to be sure, but one in which the possibilities for improving writing assessment and writing programs is very real.

3. Commentary: a challenging answer to a complex problem

What impresses me most about *Whale* is its systematic attention to the issue of social consequences of assessment programs. In doing so, *Whale* offers its readers—in a way few others have done—a constructive synthesis of work in both the fields of measurement and writing. Throughout the book, White, Elliot, and Peckham offer a vision of writing assessment design that challenges us to put concern for consequences first arguing, "If twenty-first century writing researchers attend to consequential validity to the extent that twentieth-century researchers focused on inter-rater reliability, the most important group of stakeholders, our students, would surely benefit" (p. 83). I agree. But I would argue that the significance of the point is lost in the understated nature of the claim. I have been arguing for some time (Slomp, 2008, 2011; Slomp et al., 2014) that attention to social consequences has the potential to revolutionize assessment practices.

Please cite this article in press as: Slomp, D.H. Book review. *Assessing Writing* (2016), http://dx.doi.org/10.1016/j.asw.2016.01.002

Largely speaking, the measurement community has paid little attention to consequential dimension of assessment (Lane, 2013) often to the detriment of students, schools, and educational systems. *Whale* illustrates this point. Frequently throughout the text White, Elliot, and Peckham point to the limitations of current assessment models such as the timed, impromptu assessment design so prevalent in schools around the world today. This design is problematic, they argue, because it cannot hope to possibly capture the enormous complexity of the writing construct. But does it matter if a test underrepresents the construct? Without attention to consequences, construct underrepresentation is simply a technical issue. The continued use of timed impromptu writing assessments throughout the world suggests that when framed simply as a technical issue, there is little impetus for change. However, if our field paid much greater attention to the consequences of this design for student writing development, teacher composition pedagogy, and systems of education, I predict that the impetus to reform this model of writing assessment would accelerate.

The impact of assessment design, score interpretation, and score use on the lives of real people brings into focus the enormous responsibility involved in designing and implementing large-scale and classroom assessments of writing. It also makes it far more difficult for assessment designers and users to simply dismiss the shortcomings of their assessment designs.

This attention to social consequences also highlights the importance to developing a strong program of research articulating a robust, current construct model for our field. In this respect, *Whale* challenges us to not be content with limited, dated, or post hoc construct models in the design and implementation of our assessment programs. Instead it compels us to articulate clearly developed and robust construct models as the foundations of both our instructional programs and their attendant assessment programs. This too will be an important advance for our field. As a member of the Editorial Board for *ASW*, I find that many articles I review for this journal do not clearly articulate the writing construct, nor do they demonstrate how the assessment instrument discussed in the article maps onto or reflects that construct. Almost always, this omission results in a recommendation for revisions and resubmission. This pattern demonstrates that often times in our own field we do not attend closely to the construct we are attempting to measure. I suspect that this situation reflects the influence of the psychometric community on our field. Psychometrics, in general, tends to treat construct representation as a post hoc process where—rather than a priori defining the constructs to be measured—construct features are most often inferred from statistical analysis of test-taker's results (Zumbo & Chan, 2014). The point that White, Elliot, and Peckham make in *Whale*, however, is that this post hoc approach is no longer acceptable because it does not allow for purposeful program and assessment design.

I mentioned in the introduction to this review that *Whale* does not offer simple and easy answers. It does offer, perhaps, the next best thing. Perhaps the most valuable feature of this book is that at the end of each chapter is a list of questions that can be used by readers to work through, in their own contexts, the implications of White, Elliot, and Peckham's ideas for the design and appraisal of their own assessment programs. As such, the book extends its relevance beyond its core WPA audience to both an international audience of writing teachers and to teachers of writing at all levels of education. Answering the questions posed throughout the book will help readers develop rigorous assessment programs within their own contexts.

4. Complexity: strength as weakness

Having said this, I believe that *Whale* could have offered its readers much more. In some respects the strength of its authors is also a weakness of this book. White, Elliot, and Peckham are senior scholars in the field of writing assessment. Individually, their works are among some of the most important in our field. Reading *Whale*, it is clear that these three scholars draw an enormous breadth of knowledge and experience into this book. It feels often, though, that the book is skimming the surface of this sea of knowledge. For readers who are not already well versed in the field of writing assessment, this book will feel overwhelming. Often the authors introduce their readers to a key concept, point to a body of sources in which the concept is developed more deeply and thoroughly, and then they move on. I had wished, often, that the authors would slow the text down, expand upon their ideas, and explain in greater detail their significance before moving on.

There are three areas in particular where I had wished the authors would have expanded upon their ideas.

Throughout the book, the authors emphasize the importance of robust construct representation. In Chapter 3 they present a construct model that they have derived from a range of research and policy reports. *Figure 3.1 Nomothetic span of the writing construct* and its attendant discussion (pp. 74–75) offer a valuable overview of the complex dimensions of the writing construct. The one page discussion, however, does not unpack the 39 construct elements identified. Consequently, readers are left with a broad sense of the construct, but not a detailed explication of it, one that elucidates the range of knowledge, skills, features and dispositions that collectively contribute to the writing construct. Admittedly, such a discussion would necessarily have been extensive. Had the authors taken up this challenge, though, the contribution to the field made by *Whale* would have been much more significant.

The second area where I had hoped for a more elaborated discussion was in the authors' treatment of validity theory and its history. It is clear that these authors understand validity theory in a deep, rich, and nuanced way. In *Whale*, however, their treatment of the concept is woven throughout the book. A reader, new to assessment theory will pick up pieces here and there, but will likely walk away feeling confused. For example, *Whale* interacts with Kane's (2006) IUA model of validation breezily, explaining relevant components of it as needed. For those who have not read Kane before, this treatment will be insufficient for helping them develop a coherent understanding of the concept. I remember my first time reading Kane's

Please cite this article in press as: Slomp, D.H. Book review. *Assessing Writing* (2016), http://dx.doi.org/10.1016/j.asw.2016.01.002

(2006) chapter on validation—the ordeal took the better part of a week before I fully understood the nuances of his work. Validity theory is complex, the process of validation is multifaceted. Understandings of the concept and its facets differ over time and across research communities. Had the authors of *Whale* explicated this theory—exploring in greater detail this history and range of disciplinary understanding—they would have done their readers a great service. Such a discussion would help readers who are steeped in other traditions orient themselves to the ideas presented in *Whale*.

The third area where elaboration would have been very helpful is the end of Chapter 4. In Chapters 4 and 5 the authors call for methods of data presentation that enable an integration of quantitative, qualitative, and mixed-methods data. It would have been of great value to their readers if the authors had extended Chapter 4 with a demonstration of what this might look like in action.

5. Vision for the future

In framing writing assessment as a socially situated tool used to achieve a professional community's ends, *Whale* makes clear that writing assessment has power to shape programs. In an era of accountability, where so many educators are oppressed by externally imposed assessments of often dubious quality, this call highlights the importance of teacher knowledge, and it motivates the design of high quality locally developed assessments as a tool to fight back—to reclaim professional authority and autonomy. I believe that these are important offerings for our field. As such, this book provides a vision of how to work with classroom teachers and other writing professionals to help them develop assessment programs that will have the power to transform their classrooms, their schools, their districts, and perhaps even their provincial/state and national education systems.

For those who are entering the field of writing assessment, I would recommend treating this book as a starting point, not an end. Follow the breadcrumbs laid out for you. When the authors point to additional readings, take the time to explore and understand them, before coming back to this book. The work will be difficult, but it will be well worth the effort.

References

Campbell, C. (2013). Research on teacher competency in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 71–84). Los Angeles, CA: SAGE Publications, Inc.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
Gould, S. J. (1996). *The mismeasure of man*. New York: Norton.
Hillocks, G., Jr. (2002). *The testing trap: How states writing assessments control learning*. New York: Teachers College Press.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
Lane, S. (2013). The need for a principled approach for examining indirect effects of test use. *Measurement: Interdisciplinary Research and Perspectives*, *11*(1–2), 44–46.
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.
Marcellus, K. (2014 Dec 2). It's unfair to students to decrease the weighting of Diploma Exams, says former director of provincial testing. *Edmonton Journal*,. Retrieved from: http://blogs.edmontonjournal.com/2014/12/02/its-unfair-to-students-to-decrease-the-weighting-of-diploma-exams-says-former-director-of-provincial-testing/
Melzer, D. (2015). Approaches to assessing student writing and writing programs in the age of accountability. *Writing and Pedagogy*, *7*(2/3.), 423–428.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.
O'Neill, P. (2015). A review of White, Elliot, and Peckham's very like a Whale: The assessment of writing programs. *Journal of Writing Assessment Reading List*, http://jwareadinglist.blogspot.ca/2015/11/a-review-of-white-elliot-and-peckhams.html
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.
Slomp, D., Corrigan, J., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments. *Research in the Teaching of English*, *48*(3), 276–302.
Slomp, D. (2011). Before the floor sags [Review of: Reframing writing assessment to improve teaching and learning]. *Assessing Writing*, *16*, 72–75.
Slomp, D. (2008). Harming not helping: The impact of a Canadian standardized writing assessment on curriculum and pedagogy. *Assessing Writing*, *13*, 180–200.
Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY: Springer.

David H. Slomp
University of Lethbridge, Lethbridge, Alberta, Canada
E-mail address: david.slomp@uleth.ca
Available online xxx

Please cite this article in press as: Slomp, D.H. Book review. *Assessing Writing* (2016), http://dx.doi.org/10.1016/j.asw.2016.01.002