Discussant
Coordinated Symposium
NCME 2018 Annual Meeting

Reflections: NCME 2018 Panel on Writing Analytics

Mya Poe, *Northeastern University*

1.0 Background

The utility of automated writing evaluation (AWE) for the purposes of teaching writing is one that would likely draw much skepticism from writing teachers, who most likely know of AWE through machine scoring controversies. Yet, the five papers presented at the Coordinated Symposium entitled *What Writing Analytics Can Tell Us About Broader Success Outcomes,* featured at the 2018 Annual Meeting of the National Council on Measurement in Education, raised a number of issues regarding automated writing feedback that writing teachers should find relevant.

Two questions animated the research presented by the panelists:

1. How do writing analytics derived from student writing samples relate to measures of broader outcomes?
2. How might these relationships between writing analytics and broader outcomes inform instruction and assessment to advance student learning?

The first question—i.e., the relationship between linguistic characteristics that computers can detect and scores—is one that, at face value, may look like first-generation work in AWE. Today, second-generation AWE is much broader. As the papers presented demonstrate, validity has become integral to AWE, and thus sociocultural concerns should no longer be ignored (Burstein et al., 2016.) What this shift means is that the linguistic characteristics identified must work within a larger universe of interpretation if the system is to be "smart" in helping students develop as writers. Attention therefore subsequently shifts from a focus solely on the identification of textual features to a focus on the kind of writer who is coming to the application, as well as the context in which the application is being used.

DOI: 10.37514/JWA-J.2018.2.1.14

The second question regarding the advancement of learning is also one that may be answered narrowly if the aim is unidirectional—that is, getting teachers to fall in line with computers. There is no doubt that over-testing remains a burden to U.S. students and their families. As a large-scale study by the Center for American Progress demonstrated, students take as many as 20 standardized assessments each year, with an average of 10 tests in grades 3–8; urban high school students spend 266% more time taking district-level exams than suburban students; and, in many cases, a school culture has emerged that values testing over learning (Lazarín, 2014). That burden remains, yet some of it can be lifted by the attention to formative feedback. Thus, my appeal is that second-generation AWE be developed with a clear understanding of the contexts in which it is to be used and be used in the service of improving teaching, not augmenting testing.

To address such an appeal, a multidisciplinary approach to research on AWE is needed. This multidisciplinary approach moves beyond soliciting subject matter experts' input at strategic moments in the research design to including researchers with decidedly different worldviews throughout the research process. This panel represented the possibilities of what long-term engagement with varied epistemological frameworks for understanding writing can yield. And I think the work here has important implications for what questions of fairness might look like in automated formative assessment.

In my comments, first made on April 14, 2018, and later revisited for these reflections, I want to outline three challenges that I see our colleagues wrestling with in terms of the theories of writing that are advanced in formative assessment, the goals of formative automated assessment, and the applications and consequences of such work.

2.0 Theorizing the Writing Construct

When working on multidisciplinary teams that include measurement specialists and writing studies researchers, it is important to acknowledge the different ways that the two groups theorize the construct of writing—a point made by Heather Camp in the introduction to a 2012 special issue of *Assessing Writing* on writing development and writing assessment. One perspective—a perspective that draws from linguistics—is based on syntactic maturity and thus looks to lexical and syntactical features in texts as markers for writing proficiency. For example, in "Writing Analytics and Formative Assessment," Charles MacArthur and his colleagues used Coh-Metrix "to represent four constructs based on theoretical considerations and prior research: Lexical complexity, syntactic complexity, and two types of cohesion." In "Linking Writing Analytics and Broader Cognitive and Intrapersonal Outcomes," Jill Burstein and her colleagues used vocabulary as a target because word choice tends to be a consistent and stronger predictor of external criteria, such as SAT scores.

Another perspective—a perspective brought from the field of writing and literacy studies —draws on the work of theorists, such as Lev Vygotsky and Étienne Wengner, and thus looks to rhetorical and social aims of writing. For writing and literacy researchers, the focus is on understanding writing development as a social-cultural process—a process that relies on situated participation in disciplinary contexts. As David Russell explains:

> Because writing, like problem solving, is not a *Ding an sich*, a method or technique universally applicable to all communication situations that require writing, but is instead a matter of learning to participate in some historically-situated human activity that requires some *kind(s)* of writing, it cannot be learned apart from the problems, the habits, the activities—the subject matter—of some group that found the need to write in that way to solve a problem, or carry on its activities (p. 194, emphasis in original).

According to scholars like Charles Bazerman (2007), even the reduction of genres to linguistic markers cannot ignore that genres "provide highly differentiated, scaffolded communicative spaces in which we learn the cognitive practices of specialized domains, including identification and display of relevant material, forms of inscribing experiences and data, forms of reasoning and issues to be addressed, stances to be taken, and relation to other texts in the domain" (p. 38).

For writing studies researchers, then, the unit of analysis can be quite varied, but the research always points to social implications. For example, in one study from the Citation Project, researchers studied research papers written by 174 first-year students at 16 U.S. colleges and universities. An analysis of the research papers found that students work from one or two sentences of cited sources in 94% of their citations, citing the first or second page of their sources in 70% of their citations and citing only 24% of their sources more than twice (Jamieson & Howard, 2014). While 78% of the papers include at least one incidence of paraphrase, 52% include at least one incidence of what Rebecca Moore Howard calls "patchwriting," in which writers rely heavily on the original syntax of the source material with only minimal changes. Based on their findings, Citation Project researchers concluded that "students need instruction in strategies for understanding, initiating, and entering into academic conversations and arguments" (Jamieson & Howard, 2014). In fact, much longitudinal writing research has traced the uneven development of student writers as they learn disciplinary genres through processes of mentoring and modeling. For such researchers, questions about reflection, self-efficacy, and beliefs about writing (and oneself as a writer) are central.

We see how these two viewpoints, although distinct, can nevertheless resonate within the presentations in this session. In "Linking Writing Analytics and Broader Cognitive and Intrapersonal Outcomes," Burstein and her colleagues note the mismatch in student surveys and lexical markers, suggesting that students carry a different theory of writing development—an alternative theory of writing development that we should listen to if formative feedback is to be adopted by students. Given the papers presented at the NCME conference, it seems clear that one way forward for multidisciplinary research between educational measurement and writing scholars is to focus on just what is meant by a socio-cognitive, situated view of language.

3.0 Rethinking the Goals of Assessment

Negotiating different constructs of writing is just one aspect of unpacking the complexity of multidisciplinary teams for the purposes of automated formative assessment. In doing such work, we must reimagine the goal of this kind of assessment. The goal is not solely a focus on scores for the purposes of predicting future performance but a focus on identifying markers in student writing that identify places for development of that writing. In short, we have shifted from scores to feedback. This shift is most welcome. The goal nevertheless must be extended. Guided learning, with an assumption that successfully addressing one set of issues in one piece of writing constitutes an aim, is not in-itself the endpoint. Rather, formative assessment must direct students through ever-increasingly complex dimensions of writing. The goal is not to edit. The goal is to create.

Aiming toward complexity is something that well-trained writing teachers very much understand, a point underscored by Bennett (2011), in which he argues for teachers to have strong cognitive-domain understanding: "A teacher who has weak cognitive-domain understanding is less likely to know what questions to ask of students, what to look for in their performance, what inferences to make from that performance about student knowledge, and what actions to take to adjust instruction" (p. 15). As any writing program administrator will note, good writing teachers have a strong understanding of the cognitive domain of writing. I witnessed that expertise recently during a writing program workshop at the University of Wisconsin-Milwaukee (UWM). Teachers at UWM are transitioning away from a high stakes end-of-semester portfolio assessment to a formative model of portfolio assessment that involves multiple readers and the opportunity for revision. As I worked with teachers about features of writing they value and want to retain in their new assessment model, terms surfaced such as *interaction*, *critique*, *engagement*, and *rhetorical purpose*. Teachers wanted to move students away from high school theories of writing that are often based on simplistic arguments, text length, and purple prose, and replace them with theories of writing that acknowledge the multiple tensions within a text, work closely with sources, and if possible, work toward increasingly sophisticated sentence structures that are used in the service of sophisticated arguments. To accomplish these goals, teachers provided highly-individualized feedback for each student over multiple writing tasks—feedback was buttressed by additional comments from other expert teachers. For writing teachers, then, the goals of formative assessment are realized over long periods of time.

We see this orientation towards cognitive complexity in the work of Val Ross and her colleagues in "*MyReviewers*: Understanding How Feedback Supports Writers in Higher Education." Her findings on the types and frequency of peer review responses reveal that STEM students provided more feedback related to telling as well as explaining and exemplifying, unlike their FY cohort who provided more feedback related to praising and suggesting. As this research demonstrates, to align the goal of formative assessment with the development of automated systems, researchers must come to terms with the fact that, with cognitive complexity in mind,

the goal is creation. To achieve this, strong cognitive-domain understanding must be considered as a goal of formative assessment through AWE.

4.0 Developing Evidence-Based Approaches to Fairness

Finally, the question of fairness is, of course, an issue that is generating much conversation in both the measurement and writing communities. From the educational measurement community, recent work by Zwick (2017) is especially informative. The most important considerations for our conversation at the present time are the implications of automated formative assessment:

- Have we learned our lessons from previous generations of AWE?
- What will be the washback effects of AWE on teaching?
- Who benefits from these systems?

According to Elliot (2016), "Fairness in writing assessment is defined as the identification of opportunity structures created through maximum construct representation. Constraint of the writing construct is to be tolerated only to the extent to which benefits are realized for the least advantaged." (§ 3.1) For fairness to be realized, according to Elliot, three principles must be followed:

1. *Identification of the least advantaged:* The question is not merely whether data will be disaggregated, but how we talk about the least disadvantaged—will that be based on deficit discourse or opportunity discourse? Likewise, are automated systems going to be normed to "average" students—white, able-bodied, middle-class, English-speaking students? If there is a fixation on reliability, then we will only be norming to a very small slice of students today in higher education.
2. *Maximum construct representation:* The curriculum is the place where maximum construct representation can be realized. I do not see automated feedback systems as replacing teachers, but in supporting teaching. I doubt there's a teacher in the U.S. today who would object to spell checker or grammar check being used in the writing classroom. Technologies that help students learn syntactical complexity as well as lexical complexity and other linguistic features, which are markers of more advanced writing development, are valuable. Through their use, teachers can address issues that machines are not yet good at addressing—for example, how to tell a story with your data or how to effectively synthesize viewpoints within a disciplinary conversation.
3. *Advancement of opportunity:* The potential of automated feedback systems to advance opportunity is an important conversation. For English language learners, having access to a system like Writing Mentor™ means getting *access* to writing instruction that goes beyond spell check and grammar check. It means getting access on an accessible platform—Google Docs—and getting help on the kinds of writing issues that are covered in introductory college writing courses. In the case of utility-value, note that the

> Harackiewicz et al. (2016) research showed improved course grades for underrepresented minority students (URM), especially first-generation URM students. In the research presented by Beata Beigman Klebanov and her colleagues, "Utility-Value Score: A Case Study in System Generalization for Writing Analytics," there is much to learn about the ways that writing features shift to a new institution (and, thus, a new student population), a new subject matter course, or a new variant of the original task. While that may be a limit of the reported study, it is also an important finding in terms of the deeply-situated nature of written communication. Such windows of opportunity are significant, and I encourage us not to lose sight of them.

Because of the enormous potential presented in the multidisciplinary panel to change writing instruction, I conclude my response by encouraging both the educational measurement community and the writing studies community to advance evidence-based standards for writing assessment in general, and for automated models in particular, that make fairness a central consideration.

Author Biography

Mya Poe is an Associate Professor of English and Director of the Writing Program at Northeastern University. Her scholarship has appeared in *College Composition and Communication, Journal of Business and Technical Communication, Journal of Writing Assessment,* and *Across the Disciplines*. Her co-authored and co-edited books have won the CCCC Advancement of Knowledge Award (2012) and the CCCC Outstanding Book of the Year (2014). She has also guest-edited special issues of *Research in the Teaching of English* (2014) and *College English* (2016) and is co-editor of the *Oxford Brief Guides to Writing in the Disciplines*. She is co-author, most recently, of "Civil Rights and Writing Assessment: Using the Disparate Impact Approach as a Fairness Methodology to Determine Social Impact," published in the *Journal of Writing Assessment*. With Asao B. Inoue and Norbert Elliot, she is editor of *Writing Assessment, Social Justice, and the Advancement of Opportunity*. Currently, she is working on a monograph entitled *Intended Consequences: Making Writing Assessment Fairer*.

References

Bazerman, C. (2007). Genre and cognitive development: Beyond writing to learn. SIGET [on line]. Retrieved from http://www3.unisul.br/paginas/ensino/pos/linguagem/cd/English/5i.pdf

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, *18*, 5–25.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York: Routledge.

Burstein, J., Beigman Klebanov, B., Elliot, N., & Molloy, H. (2016, September). A left turn: Automated feedback & activity generation for student writers. *Proceedings of the 3rd Language Teaching,*

Language & Technology Workshop. San Francisco, CA. Retrieved from http://www.isca-speech.org/archive/LTLT_2016/pdfs/LTLT2016_paper_2.pdf

Burstein, J., & McCaffrey, D. (2018, April). *What writing analytics can tell us about broader success outcomes*. National Council for Measurement in Education, Coordinated Session A4. New York, NY.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2018, April). *Linking writing analytics and broader cognitive and intrapersonal outcomes*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*. Retrieved from http://journalofwritingassessment.org/article.php?article=98

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

Harackiewicz, J., Canning, E., Tibbetts, Y., Priniski, S., & Hyde, J. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*, 745–765.

Jamison, S., & Howard, R. M. (2014). Citation Project Study #2 Writing from sources: Information selection, use, and citation in undergraduate research papers from sixteen US colleges and universities. Retrieved from http://www.citationproject.net/studies/writing-from-sources/

Lazarín, M. (2014). *Testing overload in America's schools*. Washington, DC: Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf

MacArthur, C., Phillipakos, Z., & Jennings, A. (2018, April). *Writing analytics and formative assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY.

Ross, V., Moxley, J., & Eubanks, D. (2018, April). *MyReviewers: Understanding how feedback supports writers in higher education.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY.

Russell, D. R. (1993). Vygotsky, Dewey, and externalism: Beyond the student/discipline dichotomy. *Journal of Advanced Composition*, *13*, 173–197.

Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions*. Cambridge, MA: Harvard University Press.