RETROSPECTIVE. IMPLICATIONS OF AUTOMATED SCORING OF WRITING

Laura Aull
University of Michigan

Popular notions of automated scoring are often oversimplified, and grim. They bring to mind product-oriented treatments of writing. They summon images of machines replacing teachers. They conjure inhumane outsourcing—an un/necessary evil, the equivalent of getting a robot on the phone who cannot understand you and leaves you desperately annunciating, *operator!*

The implications of being misunderstood are far more serious than an unsuccessful phone call, of course. Scoring algorithms are unable to parse some students' creative ideas because they are in language deemed nonstandardized by schools that reinscribe prescriptive and oppressive histories (Hammond, 2019; Perryman-Clark, 2013). Automated scoring that is most able to focus on machine-readable text does not focus on 'languaging': the mental processes of meaning-making that surround the text produced (Ivanič, 2004). In both of these examples, automated scoring belies the practices and principles most of us support in our writing courses.

Yet we cannot ignore automated scoring any more than we can ignore any widespread writing assessment approach. It is part of student writing today, and it touches on all the major themes in this collection: technical issues; evolving ideas about writing; teachers' and students' lived experiences; policy; and embedded concerns about reliability, validity, bias, and fairness. Automated scoring requires our engagement, even as it can be hard to know what to think, between media representations of automated scores, outsourced automated tools, high stakes for students and instructors, and the varied demands on writing students, educators, and administrators to use automated scoring tools.

This essay strives to offer an overall look at automated scoring in writing assessment over the past two decades, particularly how it constructs student writing and writers and how writing educators might engage with it. To do so, I draw on three articles from the *Journal of Writing Assessment* that illustrate writing educators' critical engagement with automated scoring:

1. Validity of Automated Scoring: Prologue for a Continuing Discussion of Machine Scoring Student Writing by Michael Williamson (2003)

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.23

2. Critique of Mark D. Shermis & Ben Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis" by Les C. Perelman (2013)
3. Globalizing Plagiarism & Writing Assessment: A Case Study of Turnitin by Jordan Canzonetta and Vani Kannan (2016)

As the titles suggest, the articles have different perches and points of entry vis-à-vis automated scoring. Williamson describes two stakeholder groups implicated in automated scoring, writing educators and measurement professionals, in order to explore each group's perspective and the dangers of keeping them separate. Perelman offers a data-driven appraisal of automated scoring by critiquing a foundation-funded study's unfounded claims—in turn, offering an illustration of the important possibilities of the cross-pollination called for by Williamson. Canzonetta and Kannan write about *Turnitin.com*, a plagiarism detection software, as it contends to move into global formative assessment, bringing with it particular ideologies.

Together, the articles help us consider five questions as they have been answered over time:

1. What is automated scoring?
2. What does automated scoring do?
3. What is the role of automated scoring (or, are humans good at scoring)?
4. What responsibility do writing educators have vis-à-vis automated scoring?
5. What might the future of automated scoring of writing look like?

Considering the three articles in light of these questions allows us to interrogate automated scoring, its implications, and its future possibilities for student-centered assessment.

WHAT IS AUTOMATED SCORING?

While each article focuses on various components of automated scoring, together they illustrate its expanding scope over the past two decades. In 2004, Williamson defined automated scoring as "the use of computer algorithms to simulate holistic ratings of student writing" (p. 86).[i] In this definition, Williamson's pairing of algorithms and holistic scoring, if obvious, is interesting: computer algorithms measure discrete, direct features—say, the number of words per sentence, or the frequency of certain content words—and holistic ratings are overall arguments about a piece of writing (e.g., a rating of whether the writing is successful in light of the task). Here, I use *arguments* in Kane's (2013) sense: a score is an interpretive argument about the writing. In the case of an automated score, a holistic rating for a full piece of writing is an algorithm-based interpretive argument about how the discrete features come together in a single score. The evidence for the argument includes the features the algorithm is designed to measure.

Depending on the target writing feature(s), algorithms can use a formula consistently, more easily than humans (for instance, a computer algorithm can measure lexical sophistication, per use of rare and varied content-related words, consistently), a point to which we'll return below. As these three articles suggest, the key question for automated scoring centers on what algorithms can actually identify and evaluate, and to what end. In other words, "the question is whether the task itself is computable" (Williamson, 2003, p. 96)—the extent to which writing can be analyzed and evaluated by machines.

On this, the articles show consensus overall: automated scoring cannot evaluate the full complexities of writing as situated rhetorical action, though it poses important opportunities for writing educators and assessment. Williamson argues that an automated score cannot replace a human rater score, but automated scoring can augment the teaching of writing. Perelman shows the risks involved when automated scoring tools are "judged like the answer to a math problem or GPS directions" (para. 6); he also argues that automated scoring demands rigorous statistical analysis and offers important information. Canzonetta and Kannan likewise note some limitations of automated scoring tools and call for critical engagement, particularly with the "global cultural work" of plagiarism detection tools.

As we can see, the three articles imply a broader definition of automated scoring, as the use of algorithms to evaluate various aspects of writing, whether or not the expectation is to simulate human scorers. For those cases in which an automated scoring tool alone is used to determine a writing score (e.g., ACCUPLACER, WritePlacer, and WritePlacer ESL), Williamson's early definition still applies. In all cases, we can think of automated scoring as an approach that uses algorithm-based evaluations of writing as evidence for an interpretive argument about said writing.

WHAT DOES AUTOMATED SCORING DO?

Embedded in our definition above is that automated scoring algorithms determine what counts, and doesn't count, in a piece of writing. An overall score based on a given algorithm draws inferences from those aspects determined to count. In so doing, any given automated tool constructs writing, assessment, and writers in particular ways.

Constructing Writing

Computer algorithms are written by humans and carried out by computers; they are limited to (and enhanced by, depending on your perspective) what computers can do. Any given automated scoring algorithm implies what matters most, constructing writing according to what it measures, such as mechanical choices, length (Perelman), text-matching, and/ or standardized written academic English (Canzonetta and Kannan). In turn, it implies that certain aspects of writing *don't count*—the things not measured, if the automated score is the only score used.

All the articles emphasize that like any kind of writing, academic writing is an integration of many processes, and this is part of concerns about automated scoring. Using the same automated scoring tool on each one—or comparing automated scores across them—belies the situated rhetorical action entailed in each one. The articles underscore that automated scoring does not construct writing as situated language use: it cannot account for writing as rhetorical action (Perelman), writing as situated literacy (Williamson), and writing and source use as culturally-specific action (Canzonetta & Kannan). The use of different automated tools, in turn, emphasizes different conceptions of assessment validity, which is also always situated (Williamson, 2003).

Constructing writing through design and use of automated scoring can also point to the lack of a clear writing construct. This is a problem Perelman delineates in his critique of the Shermis and Hamner study: "Without [any explicit construct of *writing*], it is, of course, impossible to judge the validity of any measurement" (para. 6). An illustration Perelman offers is the use of multiple constructed response tasks compared in the same way, e.g., some that require understanding and incorporation of included reading texts, and some that do not.

Constructing Assessment

Williamson traces ideas about validity with attention to the situated nature of literacy. Williamson's attention to the "complexity of validity inquiry" (p. 259) illuminates differences in conceptions of assessment "between English Studies and educational measurement, the difference between social science and humanistic disciplines" (p. 260). To date in 2004, Williamson argued, many researchers in English Studies subscribed to "an older notion of validity," thereby "unwittingly missing" more contextualized, less rigid conceptions of validity for writing assessment (p. 262). In other words, early questions about validity focused on a given test, and whether it measured what it purported to measure. In those cases, assessment is constructed as a process of consistent measurement, regardless, for instance, of the validity of the writing construct, or the particular abilities emphasized in an assessment task.

Assessment tools construct writing assessment through what they do *not* evaluate as well. Deane et al., all ETS researchers, (2013) show that "[automated scoring] systems do not explicitly evaluate the validity of reasoning, the strength of evidence, or the accuracy of information" (para. 2). They illustrate how this poses a risk because it can mean a disconnect between what assessors value and what an automated tool can measure. In such cases, the scoring makes an interpretive argument about a piece of writing based only on partial evidence about the piece of writing.

Perelman shows that the *scoring* part of automated scoring is important not only in what it interprets but also in how it is represented. His article reviews a study that represents its findings as though they suggest that automated scorers are as accurate as human raters. Yet as Perelman shows, the automated scores were rounded to integer values in ways that favored the automated scores. Perelman writers, "Essays scores, be they holistic, trait, or analytical, always are continuous variables, not discrete variables (integers), even though graders almost always have to give integer values as scores" (para. 14). Interrogating automated scoring and representations thereof have high stakes for how assessment is constructed vis-à-vis policy decisions. The paper Perelman critiques, for instance, was sent to the Partnership for Assessment of Readiness of College and Careers and the Smarter Balanced Assessment Consortium. Thus Perelman argues that automated scoring demands rigorous statistical analysis and offers important information in light of potential policy decisions.

Constructing Writers and Assessors

While the first three articles highlight how automated scoring constructs writing and its measurement, Canzonetta and Kannan's article more specifically illuminates how writers are constructed by automated scoring. They question, "How is the student plagiarist being discursively constructed? What are the implications of these constructions as Turnitin rolls out its assessment platform?" (p. 298). Their answers to these questions show how Turnitin.com's strategies construct those who assess writing, too.

Canzonetta and Kannan discuss "three primary rhetorical strategies for advancing Turnitin" that construct writers and assessors in culturally-specific, hierarchical ways. They include: (1) plagiarism detection as social improvement that forms modernized, idealized, western students; (2) plagiarism as a national concern with ramifications for citizenship, economy, and character; and (3) plagiarism detection as requiring standardized, western policies through private/

public partnerships (para. 11). These strategies construct writing as rule mastery, writers as needing to be regulated, and instructor-assessors as regulators.

In sum, the articles remind us that vis-à-vis the question *what does automated scoring do?*, we can answer: it constructs writing, writers, and assessors in particular ways. As a writing assessment tool, any automated scoring tool is constitutive; all assessment activities are complex and value-laden activities, whether baked into an algorithm or required of a human (Aull, 2017). Thus the authors remind us that, as is any writing assessment approach, automated scoring is an opportunity to see what we prioritize in writers and writing—a chance to understand and question what that is and what's left out.

WHAT IS THE ROLE OF AUTOMATED SCORING (OR, ARE HUMANS GOOD AT SCORING?)

Underpinning questions about automated and human scoring is the fundamental question *what is "good" scoring?* In early writing assessment research especially, good scoring was often described in terms of reliability, in other words, consistent scoring. It was also often described as scoring with agreement between scores (human, automated, or both). We will consider agreement and reliability below, but a broader answer is that even pondering "good" automated and human scoring entails more questions posed by the three articles: questions about the widespread use of automated scoring and about the strengths and limitations of both machine and human scoring.

The Inevitability of Automated Scoring

"Two things are certain," writes Williamson. "One, automated scoring programs can replicate scores for a particular reading of student writing, and this technology is reliable, efficient, fast, and cheap. Two, automated scoring has been and will continue to be used in various large-scale assessments of student writing" (p. 256). Canzonetta and Kannan describe how Turnitin.com frames automation in similar ways, emphasizing "efficiency, adaptability, and individual choice" (p. 305).

We cannot escape from this answer: automated scoring is widespread. It is used in large part because it is an inexpensive way to assess student writing, which is otherwise labor- and time-intensive. That, and a longer history of testing dating back to the early 20th century made writing a skillset for ranking individuals against one another in a single test (Elliot, 2005; Hammond, 2019; Aull, 2024). At the same time, the three *Journal of Writing Assessment* articles help frame automated scoring not only as a large-scale inevitability, but also as a reminder that we should not accept any assessment approach without ongoing questions about its limitations.

Automated Scoring Limitations, Human Strengths

Computers cannot interpret or infer; they can only analyze the observable components of a text, which are based on the preferences and designs of the developers or the graders whose models were used. Thus there are always limitations—in the basic example above, essay length is easy for automated scoring tools to detect, but it is not always a sign of valued written detail. Perelman also reminds us that automated scoring tools can be "gamed"; for instance, writers can use conventionally prioritized academic writing choices, including cohesive ties ("however,") and large words (lexical complexity). They can likewise be gamed by *avoiding* choices conventionally devalued in standardized usage preferences, including "sentence fragments" and sentences beginning with conjunctions like *and*. None of these choices necessarily have to do with genre-specific idea development. But they could lead to a positive automated score because automated tools can directly measure these features. Deane et al. (2003) describe how automated scoring systems rely "on measures of such things as the structure and elaboration of student essays, the sophistication of vocabulary, or the number of errors in grammar, usage, mechanics or style" (para. 2), defined according to the scoring system designers.

By contrast, automated tools cannot look at writing as a situated act, contingent on audience, genre, and purpose of a specific rhetorical situation. In the case of automated plagiarism detection, Canzonetta and Kannan show, *Turnitin.com* developers talk about writing in a way that "displaces composition and rhetoric's arguments for situated pedagogical approaches" (p. 305). Perelman shows how automated tool design can elide common expectations for writing such as coherence, e.g., by not accounting for paragraph breaks.

Alternatively, human raters are capable of assessing writing tasks as situated actions. Deane et al. (2013) show that human raters can attend to genre-specific critical thinking on complex writing tasks involving source text integration in ways that automated scoring cannot. In the case of exams purportedly measuring students' preparation for college writing, construct validity, or the relevance of the constructed response task, depends on having human raters. Human raters can consider how writers have engaged with source text use, a common expectation of college writing.

Indeed, human scorers can read for audience-specific choices, idea development with or without explicit transitional phrases, and integration of others' ideas. These practices support today's sociocultural conceptualizations of writing as rhetorical action among language users in a specific text and community. As Perelman puts it, "Writing is foremost a rhetorical act, the transfer of information, feelings, and opinions from one mind to another mind . . . The essence of writing, like all human communication, is not that it is true or false, correct or incorrect, but that it is an action, that it does something in the world" (para. 4). Humans more easily read and write in this sociocultural way than machines. Yet Perelman's point reminds us that human and automated scorers alike can belie this sociocultural conceptualization when they focus on error-hunting rather than situated meaning-making.

It follows that like writing and literacy, writing assessment and its validity are situated. We can see in Williamson's 2003 article that ideas about validity needed expanding at the time: he cautions against seeing validity only as reliability. In other words, he decries the notion that validity means "a test has to measure what it purports to measure," because validity is situated "in a particular use of a test, in a particular context, at a particular time" (p. 267). This emphasis on use anticipates Kane's work on interpretation and use arguments that are part of any test: scores represent inferences drawn from assessments. Those scores (and use of those scores, such as admissions or course placement) make interpretive arguments (Kane, 2013). Likewise, recent notions of reliability have expanded, calling into question prevailing standards built on narrow testing constructs, moving instead toward reliability based on the measurement, conditions, and objectives of complex writing performances (Ross & LeGrand, 2017).

Human Limitations, Automated Scoring Strengths

The last section makes it easy to see why it is common to assume that human scoring is ideal—that, to use an example from the opening, automated scoring is at best a necessary evil. Indeed, trust in humans underpins the prevailing practice of training an automated scoring model against human scores. Canzonetta and Kannan write that Turnitin's "intelligent assessment" alleges to grade papers like humans. More generally, Williamson calls for assurance that "the goals of people . . . drive the development of automation, not the automation itself" (p. 272).

But as writing and research remind us, humans also have limitations; like automated scoring, human scoring merits critical investigation. Human scorers working in the U.S., Canzonetta and Kannan emphasize, have overwhelmingly inherited "a culture of standardized testing" and "hegemonic cultural expectations about writing and authorship." In particular, humans have subjective associations with usage choices and with particular constructed response tasks (Johnson & VanBrackle, 2012; Perryman-Clark, 2016). Most U.S. language arts and writing instructors have taught rules based on 18th century usage preferences rather than "what language is and allows human beings to do" (Gere et al., 2021; Smitherman, 2017, p. 6). They have learned to "know good writing when they see it" rather than to analyze language patterns (Aull, 2021; Lea & Street, 1998). Research shows that when presented with an expression of non-standard English, a typical rater will undervalue the essay even though an answer may be functionally equivalent to a response given in standard English (Shermis, Burstein, Higgins, & Zechner, 2010, p. 5). It is hard to train human scorers to consider language diversity without bias, even though research shows that alternatives to conventional structural choices, for instance, vis-à-vis articles, verb morphology, nouns, and verbs, do not inhibit meaning or cause lower scores for many human raters (Allen, Crossley, Kyle, & McNamara, 2014).

Research also shows lack of agreement between humans in terms of what matters most and when. Studies show that style errors, for instance, are context dependent, and agreement on when a style error occurs may differ from person to person (Crossley, Bradfield, & Bustamante, 2019). The writing task can influence not only human scores, but also the extent to which humans agree on scores.

Automated scoring technology, then, can be one way to strive to mitigate socially-constructed bias against nonstandardized usage preferences, making it easier for a blueprint to read for ideas expressed in diverse ways. Shermis et al. describe that automated scoring tools would have the capacity to overcome this human limitation if the relevant affected variables associated with nonstandard English can be isolated and adjusted (Shermis, Burstein, Higgins, & Zechner, 2010).

Human and automated scoring together. Agreement between human and automated scoring is much discussed in research on automated scoring methods. Yet the three articles suggest that the more important question is, instead, what each can read and interpret. According to Williamson, different approaches to agreement highlight different disciplinary approaches of writing educators versus education measurement professionals. While for an assessment professional trained as a social scientist, the "immediate question is whether the procedures used by automated scoring engines simulate the scoring process of human raters," a writing educator trained in English studies expects that literacy entails various readings on one text (Williamson, p. 265). One way to read this tension is as a productive one. With a view of writing as complex written action contingent on rhetorical situations, what aspects of academic writing are rigid, and what expectations need not be?

In a critique of the Shermis and Hamner study, Perelman shows that exploring agreement includes important questions related to constructed response tasks, resolving disagreement, and evaluation criteria. On the first point, Perelman shows that constructed response task influenced agreement: human scorer agreement in response to one essay task (essay set 2) was stronger relative to other essay set scores, and relative to machine scores. On the second point, Perelman shows that use of integer scores necessarily belies the fact that writing scores are continuous. In human scoring, resolving scores involves adjudicators rather than rounding; in the Shermis and Hamner study, Perelman shows how scoring resolution procedures were used in ways that privileged machine scorers and penalized human readers. Finally, Perelman shows that in certain evaluation categories, humans agreed more than automated scorers: those categories which measured ideas, content, organization, style, and voice, had an exact agreement value of 0.76, compared to the range of machine values of 0.55-0.70. All of the above provide support for Perelman's argument that in contrast to the widely-reported Shermis and Hamner claims about the reliability of machine scoring, actually, "the data provide some, although not conclusive, support for the assertion that human scorers performed more reliably than the machines, especially on longer papers that were scored for writing ability rather than solely on content" (para. 3). Perelman's work suggests another implication for calibrating for agreement between automated scoring and human raters is that, while automated scoring may be developed to imitate human assessors, human raters, too, become calibrated to evaluate like automated scoring tools.

Canzonetta and Kannan stress that a role of automated scoring is imparting culturally-specific ideas: there is no global, human agreement on a definition of plagiarism and fair use, thus any automated plagiarism tool imposes its moral and pedagogical conceptualizations on its users. If we connect this point to the other articles, we can consider the extent to which agreement itself is a culturally-specific value—a value situated in field, place, and time. In the case of Turnitin.com, Canzonetta and Kannan describe, it is a Western view of character, citizenship, and economics that is imposed. Under the guise of plagiarism detection, they write, these views are imposed but under-acknowledged. Canzonetta and Kannan call for critical attention to the views underpinning scoring and cases of automated and human scoring dis/agreement. Their call poses important considerations for culturally-specific ideas in solely human scoring as well.

Together, all of these ideas point to *a* role, but not *the only* role, for automated scoring in writing assessment. Automated scoring may provide a fuller picture of performance in complex assessment tasks that involve both reading and writing. Perelman's article underscores that the reliability of humans and of automated scoring is a site for investigation. While evidence shows that humans can read with genre-specific attention, evidence also shows they can read with socially-constructed bias against certain kinds of language use. Thus one way to respond to the questions *what is the role of automated scoring?* and *are humans good at scoring?* is to say that any scoring, and how it is represented, merit critical, ongoing attention from writing educators and educational measurement professionals alike.

WHAT RESPONSIBILITY DO WRITING EDUCATORS HAVE VIS-À-VIS AUTOMATED SCORING?

All three articles illuminate the risks of keeping writing and measurement specialists apart. First, Williamson notes a lack of cross-talk between "writing teachers" and the "assessment community," the two stakeholder groups most implicated in the scoring of writing. He explores "beliefs and assumptions held by each side" (p. 254), including disciplinary differences entailing different goals: more humanistic approaches emphasizing social context in college writing studies; more scientific approaches emphasizing aggregated patterns for assessment professionals.

Separation between the stakeholder groups, Williamson notes, means a lack of "productively learn[ing] to talk to each other about automated scoring." It means insufficient questioning of key assessment concepts such as validity, which becomes a "taken-for-granted ubiquity" in the lived experiences of students and teachers by becoming a "totalizing" and "naturalized concept and invisible instrument of rigor." Describing the early 21st century, Williamson writes, "English teacher response to automated scoring has been limited and . . . does not refer to any of the evidence presented by the developers of automated scoring programs" (p. 254).

Alternatively, Williamson describes, debate and understanding across the groups can challenge those in English Studies to address basic procedural issues of social science with questions of validity. It can challenge those in social science to consider validity as a situated construct, one that must observe the same situatedness that literacy theorists have been articulating for some time. Ultimately, the groups can come together around "the shared goal of moving toward more reliable and efficient ways to measure educational achievement and writing ability" (p. 254). At minimum, Williamson writes, "automated scoring is an incredible research opportunity through which we can explore the many different ways student writing can be read, valued, and sanctioned" (p. 254).

All three articles are examples of how productive examination of automated scoring is facilitated by engaging educational measurement and writing education. Perelman's article exemplifies the possibility of multiple methods and perspectives coming together. He uses statistical tests to expose unsupported claims in Shermis and Hamner's paper in ways valued by social scientists, and he attends to constructed response task as a situated written action in ways valued by those in English studies. Perelman's critique also shows the importance of assessment and writing researchers' engagement with publicly-rendered claims about writing and scoring. Shermis and Hamner's study findings, critiqued by Perelman, circulated widely: following Shermis's presentation of their findings at the National Council on Measurement in Education's annual meeting, the study was cited in *Inside Higher Ed* and *The New Scientist,* and by a press release from the University of Akron.

Canzonetta and Kannan add additional emphasis, calling for cross-talk about the role of large corporations in global automated assessment services. In so doing, they take up Grabill's call for more attention to automated writing technologies by composition and rhetoric scholars, stressing that "[g]lobally, millions of students are subjected to writing technologies that writing experts did not design" (p. 296). Canzonetta and Kannan specifically outline plagiarism detection tools (PDSs), using Turnitin.com's success as an example of corporate influence in U.S. universities. They analyze how PDSs constitute instructors (presumed to be members of the "Turnitin.com educational community") as preservers of ethical and moral standards, positioned antagonistically against students, and assumed to be consistent across institutions and geographic locations. They call for direct engagement, so that there is greater understanding of the global cultural work of automated plagiarism and assessment tools.

In sum, the articles underscore that writing specialists have a responsibility to engage critically with automated scoring. The alternative, they imply, constitutes risks and missed opportunities. They bring to mind White's urging: "Assessment of writing can be a blessing or a curse, a friend or a foe, an important support for our work as teachers or a major impediment to what we need to do for our students. Like nuclear power, say, or capitalism, it offers enormous possibilities for good or ill, and, furthermore, it often shows both its benign and destructive faces at the same time" (White, 1994, p. 137). In none of these metaphors is there an option for writing educators to leave alone automated scoring as part of writing assessment.

WHAT MIGHT THE FUTURE LOOK LIKE?

Like all assessment technologies, automated scoring of writing highlights and precludes particular constructs of writing and beliefs about writers. These can be implicit and taken for granted, particularly when they are established, normalized, and widespread. Alternatively, these articles make automated scoring a site of critical engagement in ways that expose implicit ideas. They help us question: What kinds of writing and writers are valued in automated scoring? What kinds of writing and writers are valued in human scoring? What can we learn from each one?

This kind of critical engagement helps us, in Williamson's words, "study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm" (p. 270). With one more article in the *Journal of Writing Assessment*, Ellen Cushman's (2016) "Decolonizing Validity," they pave the way for more inquiry-oriented, student-centered assessment. To close, I consider related possibilities for automated scoring, for making it a site of investigating rather than damning language difference, a site of collective exploration rather than a site of top-down design and individual mastery.

Decolonizing Validity

The articles by Williamson, Perelman, and Canzonetta and Kannan conceptualize automated scoring as a site for ongoing inquiry into available tools and decisions about their use. They call for understanding writing, writing assessment, and validity as situated in rhetorical situations. They show how the use of automated scoring tools can entail a cyclical approach to validity: if we define validity narrowly and measure it in narrow tests, we learn only about those narrow conceptualizations. If validity is a matter of narrowly-defined, consistent scoring, then an automated scorer can measure length, mechanics, and lexical cohesion in a timed writing task, for instance, and be valid. Alternatively, if validity is instead a matter of fairness defined as equitable distribution of scores across different student groups, then a valid test must do very different things.

Cushman argues that to date, the "concept of validity creates the colonial difference as it maintains social, epistemic, and linguistic hierarchies." It does so by "identifying what is objective and what is evidence," and by hiding its social construction: validity "is a naturalized concept and invisible instrument of rigor that totalizes the realities of students and researchers" (para. 7). Drawing on Tiostanova and Mignolo's (2012) phrasing ("dwelling in the borders" in order to '"change the construct" itself), Cushman offers an alternative conception of validity, one in which "[d]welling in the borders begins with the knowledge, languages, histories, and practices understood and valued by the people who live these realities" (para. 25). In this conceptualization, validity is collectively constructed and navigated.

In turn, validity evidence tools work "not as a way to maintain, protect, conform to, confirm, and authorize the current systems of assessment and knowledge making, but rather as a way to better understand difference in and on its own terms." In other words, validity could be seen not as a way to hold individuals to one set of metrics determined by an external group—not as "a way to maintain, protect, conform to, confirm, and authorize the current systems of assessment and knowledge making." Rather, validity could be seen in terms of descriptive power: what it helps us learn about difference. In this approach, validity measures do not "mak[e] [one] experience into a universal one, the baseline against which all Others are tested and their knowledges and languages are deemed deficit to" (para. 26). A valid assessment approach would thus be one that "seek[s] to identify understandings in and on the terms of the peoples who experience them" (para. 26).

These ideas, in turn, carry implications for students' learning alluded to by Perelman, Williamson. They carry implications for linguistic diversity alluded to by Perelman and Canzonetta and Kannan. To unpack these ideas, let's briefly consider what they entail in terms of exploring difference, formative assessment, and critical language analysis.

Exploring Difference

Williamson underscores that "fluent adult reading" expects different views from different readers (Williamson p. 265). Generally, automated scoring and inter-rater agreement expect the opposite: they expect "convergent reading." Likewise, Perelman demonstrates the dangers of disparate approaches to resolving difference. And like other studies focused on agreement between human scorers, and/or between human and machine scoring, Deane et al. prioritize what Williamson calls convergence: the smaller the difference, the better (and "ideal for operational use" are very small differences in scores inferred from reading).

Here we see a good example of Cushman's point: in this case, disparate reading is, in a sense, invalid; validity rests on agreement in reading and inferences. What if we could construct writing, and reading one's own and other's writing, not as a site of deciding whether it was right or wrong, but of exploring the inferences drawn from machines and from humans? What would it take to create that world? Canzonetta and Kannan underscore that this is not easy, because rhetorics of standardization and consistency are beneficial for Turnitin's business model.

Formative Assessment

Canzonetta and Kannan describe that "Formative assessment necessitates that teachers respond to students' needs, personalities, struggles, and strengths; and get to know them apart from their writing." Ultimately, Canzonetta and Kannan caution against automated plagiarism tools' role in formative assessment, but they do point to that role as a site for critical investigation: "it is important to critically interrogate Turnitin's rhetorics of formative assessment, which obscure the company's cooptation of student data and potential to undermine writing program goals" (para. 27). This is all the more important because the message from plagiarism tools can promote formative uses that ultimately "aim to quell critique and breed a compliant, submissive population of students" instead of a more student-centered invitation of the students' active questioning of ideas about plagiarism and writing.

One way that automated scoring could be used would be in formative reports for students' use. Following Cushman, these reports could be a site of exploring difference. What range of responses were there? What did these differences achieve? How did they respond to or change the rhetorical situation of the task?

Critical Language Analysis

Cushman writes that "Validity is on the one hand instrumental tool, which was established to manage peoples, knowledges, lands, governments, and institutions, and on the other hand, a meta-discourse which reified the social, linguistic, and epistemological hierarchies that made it possible, hence further securing its own position of authority to identify what counts as valid" (para. 11). Any use of automated scoring framed as evaluating whether writing is "good" is an example of such metadiscourse. Alternatively, framing language patterns as situated opens space for existing and valuable language diversity.

In other words, the use of automated tools and human reading can be used to identify and explore patterns descriptively, if only we can frame difference this way. Descriptive labels are possible when we chart linguistic patterns. For instance, language can be formal and informational, with many noun and prepositional phrases, such as academic writing. Language can be informal and interpersonal, with many verbs and pronouns, such as informal internet writing. This kind of charting of micro-linguistic features beyond mechanics could support Critical Language Awareness pedagogy illustrated in years of studies in student writing (Fairclough, 2014; Sanchez & Paulson, 2008; Shapiro, 2022)—pedagogy that supports students' exploration of "the social and linguistic rules" of their own language use (Smitherman, 2017, p. 6). Williamson alludes to the opportunities for automated tools to chart language patterns, noting the emerging developments in natural language processing.

Research in writing studies and assessment offers examples of how automated tools might support these efforts. Based on corpus linguistic analysis, research tests popular writing advice offered to students, showing that it doesn't always bear out (Lancaster, 2016). Similar studies map linguistic features in order that students can use linguistic patterns in analysis of their own writing (Aull, 2015, 2020). Other research questions, more broadly, the usefulness of automated tools to map out student writing features (Crossley, Kyle, & McNamara, 2015).

CONCLUSION

In the sections above, we've seen automated scoring framed as the use of algorithms to evaluate aspects of writing and as a site for exploring ideas about writing and writers embedded in any given approach to writing assessment. We've seen that automated scoring constructs writing, writers, and the practices of assessment in particular ways. We've seen the limitations of automated scoring and of human raters. All of these ideas point to the value of integrated discussions that make both human and automated scoring a site for ongoing, critical attention. Keeping automated scoring as a site of inquiry opens possibilities for mapping writing difference not for the sake of ranking but for greater understanding and exploring.

[i] To define "holistic ratings," we can turn to Haswell and Elliot: "the use of a scale to assign a single value mark to a whole essay and not separately to separate aspects of the essay, with scorers trying to apply the scale consistently, and with the final score for each essay derived from two or more independent ratings" (p. 1). Williamson also notes studies that established that holistic scoring is a limited form of reading.

REFERENCES

Allen, L., Crossley, S., Kyle, K., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. *Grantee Submission*. [Paper Presentation] International Conference on Educational Data Mining.

Aull, L. L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy*: Springer.

Aull, L. L. (2017). Tools and tech: A new forum. *Assessing Writing*. *33*, A2-A7.

Aull, L. L. (2020). *How students write: A linguistic analysis*. MLA.

Aull, L. L. (2021). What is "Good Writing?": Metadiscourse as civil discourse. *Journal of Teaching Writing 36*(1), 37-60.

Aull, L. L. (2024). *You can't write that . . . 8 myths about correct writing.* Cambridge University Press.

Canzonetta, J., & Kannan, V. (2016). Globalizing plagiarism & writing assessment: a case study of Turnitin. *The Journal of Writing Assessment, 9*(2). https://escholarship.org/uc/item/5vq519dr

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To Aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment, 8*(1). https://escholarship.org/uc/item/1f21q8ck

Cushman, E. (2016). Decolonizing validity. *Journal of Writing Assessment, 9*(1). https://escholarship.org/uc/item/0xh7v6fb

Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment, 6*(1), 40-56. https://escholarship.org/uc/item/3nf6r4kv

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Fairclough, N. (2014). *Critical language awareness*: Routledge.

Gere, A. R., Curzan, A., Hammond, J., Hughes, S., Li, R., Moos, A., Smith, K., Van Zanen, K., Wheeler, K., & Zanders, C. J. (2021). Communal justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication, 72*(3), 384-412.

Grabill J. (2016) Do we learn best together or alone? Your life with robots. *Computers & writing conference, [Keynote address]* Rochester, New York. http://elireview.com/2016/05/24/grabill-cw-keynote

Hammond, J. (2019). *Composing progress in the United States: Race science, social justice, and the rhetorics of writing assessment, 1845-1859.* [Doctoral Dissertation, University of Michigan].

Ivanič, R. (2004). Discourses of writing and learning to write. *Language and Education, 18*(3), 220-245.

Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing, 17*(1), 35-54.

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

Lancaster, Z. (2016). Do academics really write this way? A Corpus investigation of moves and templates in "they say/I say". *College Composition and Communication, 67*(3), 437-464.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172. https://doi.org/10.1080/03075079812331380364

Perelman, L. C. (2013). Critique of Mark D. Shermis & Ben Hammer, "Contrasting state-of-the-art automated scoring of essays: Analysis." *Journal of Writing Assessment, 6*(1). https://escholarship.org/uc/item/7qh108bw

Perryman-Clark, S. M. (2013). African American language, rhetoric, and students' writing: New directions for SRTOL. *College Composition and Communication*, *64*(3), 469-495, https://www.jstor.org/stable/43490767

Perryman-Clark, S. M. (2016). Who we are (n't) assessing: Racializing language and writing assessment in writing program administration. *College English, 79*(2), 206-211.

Ross, V., & LeGrand, R. (2017). Assessing writing constructs: Toward an expanded view of inter-reader reliability. *Journal of Writing Analytics, 1*, 227-257. https://doi.org/10.37514/JWA-J.2017.1.1.09

Sanchez, D. M., & Paulson, E. J. (2008). Critical language awareness and learners in college. *Teaching English in the Two-Year College, 36*(2), 164-176.

Shapiro, S. (2022). *Cultivating critical language awareness in the writing classroom*. Routledge.

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education, 4*(1), 20-26.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed.*)* (pp. 20-26). Elsevier.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*: Routledge.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 213-246). Routledge.

Smitherman, G. (2017). Raciolinguistics,"mis-education," and language arts teaching in the 21st century. *Language Arts Journal of Michigan, 32*(2), 4-12.

Tiostanova, M. V., & Mignolo, W. (2012). *Learning to unlearn: Decolonial reflections from Eurasia and the Americas*. Ohio State University Press.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance. Revised and expanded.* Jossey-Bass.

Williamson, M. M. (2003). Validity of automated scoring: Prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment, 1*(2), 85-104. https://escholarship.org/uc/item/8nv3w3w8